# Statistical Dissection of Pharmacogenetics with Machine Learning

Verbelen, Moira Inez

*Awarding institution:*
King's College London

# Statistical Dissection of Pharmacogenetics with Machine Learning

**Moira Verbelen**

MRC Social, Genetic & Developmental Psychiatry Centre

Institute of Psychiatry, Psychology and Neuroscience

King's College London

Submitted for the degree of Doctor of Philosophy

# Table of Contents

Table of Contents ..... 2
Acknowledgements ..... 4
Statement of Authorship ..... 5
Abstract ..... 6
1. Introduction ..... 8
1.1. Overview of pharmacogenetics ..... 8
1.2. Research methods in pharmacogenetics ..... 17
1.3. Pharmacogenetics in psychiatry ..... 19
1.4. Machine learning for predictive modelling ..... 23
1.5. Outline of thesis ..... 26

**Part 1: Adoption of pharmacogenetic testing in clinical practice** ..... **31**
2. Establishing the characteristics of an effective pharmacogenetic test for clozapine-induced agranulocytosis ..... 32
3. How close are we to a pharmacogenomic test for clozapine-induced agranulocytosis? ..... 38
4. Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? ..... 41
4.1. Supplementary material ..... 49

**Part 2: Machine learning prediction algorithms applied to genetic and gene expression studies** ..... **57**
5. Statistical methods ..... 58
5.1. Notation ..... 58
5.2. Statistics versus machine learning ..... 60
5.3. Traditional statistical methods ..... 61
5.4. Machine learning ..... 65
5.5. Deep learning ..... 91
5.6. Summary ..... 97
6. Diagnostic classification using machine learning and deep learning applied to brain gene expression data ..... 98
6.1. Introduction ..... 98
6.2. Methods ..... 99
6.3. Results ..... 107
6.4. Discussion ..... 115
7. Machine learning algorithms for pharmacogenetic prediction in an anti-diabetic clinical trial ..... 119
7.1. Introduction ..... 119
7.2. Methods ..... 121

7.3. Results ..... 129
7.4. Discussion ..... 148
8. Analysis of pharmacogenetic studies: comparing traditional statistical inference with machine learning ..... 151
8.1. Introduction ..... 151
8.2. Methods ..... 152
8.3. Results ..... 163
8.4. Discussion ..... 177
9. Discussion ..... 181
9.1. Summary of thesis research ..... 181
9.2. Statistical methods for pharmacogenetic research ..... 181
9.3. Challenges to pharmacogenetic research and implementation ..... 185
9.4. Future perspectives ..... 188
References ..... 191

# Acknowledgements

Firstly, I would like to thank my supervisor Professor Cathryn Lewis for her continuous support, advice, encouragement and mentoring. Her guidance has helped make the past four years a transformative, gratifying and outright extraordinary experience for me. I'm also grateful to Dr Mike Weale and Dr Raquel Iniesta, who played an important role as second supervisors, advising me with their extensive knowledge and expertise.

Secondly, I thank Professor David Collier and Dr Karim Malki for kindly welcoming me to their team at Eli Lilly. The industry placement was greatly informative and genuinely enjoyable, hence the repeated extensions. I am also grateful to everyone at Eli Lilly who shared their datasets with me.

Furthermore, I thank the Medical Research Council and Eli Lilly and Company Ltd for funding my PhD.

Finally, special thanks go to my husband Edward, my sister Viola and my parents, Inge and Jean-Pierre, who inspired me to do this PhD in the first place and who were immensely supportive and always there for me.

# Statement of Authorship

Throughout this thesis, the scientific convention of using the pronoun *we* when describing the research was followed. However, all work in this thesis was performed and written by Moira Verbelen, with the exception of the following:

Chapter 6:

- The gene expression data were collected, cleaned, pre-processed and quality controlled by others prior to this analysis.

Chapter 7:

- The data were collected in a clinical trial sponsored by Eli Lilly and cleaned by others prior to this analysis.

Chapter 8:

- The data were collected in a clinical trial sponsored by Eli Lilly and cleaned by others prior to this analysis.

In addition, supervisors and co-authors of the papers included in this thesis offered external advice on the analyses, interpretation of results and editing of the manuscripts.

## Abstract

A major challenge for personalized medicine is to identify biomarkers that predict response to therapeutics. Current experience indicates that few pharmacogenetic biomarkers are individually predictive, since such biomarkers usually lack the sensitivity and specificity to achieve a clinically meaningful prediction on their own. The future is in combining multiple biomarkers together with clinical and other characteristics, with the aim of producing multivariable prediction algorithms that can serve as decision support tools for personalizing medicine. This PhD project approaches pharmacogenetics from two different angles. Firstly, we investigate specific hurdles to the adoption of genetic tests to guide pharmaceutical treatment. We study the characteristics of a pharmacogenetic test that predicts the development of a serious adverse drug reaction caused by the antipsychotic clozapine and model the requirements for the test to be clinically useful. In addition, we assess the cost-effectiveness of pharmacogenetic testing by means of a literature review and estimate how this would change in a future where genetic information is available at no additional cost at the time of prescribing, for example via an electronic health record. The second emphasis of this PhD is on the application of machine learning methods to predict pharmacogenetic responses in Phase 2 clinical trials. Genetic studies are highly dimensional and traditionally genetic variants are investigated independently in univariate analyses. Alternatively, machine learning algorithms can be used to model large numbers of variables simultaneously, even if these variables are correlated as is the case for genetic variants. These methods optimize predictive ability and a wide range of linear and non-linear algorithms exists. Two clinical trials and one gene expression case-control study in different disease areas were analysed using machine learning (elastic net, random forest, support vector machine) and deep learning (neural network) methods with the aim of predicting efficacy and safety measures using genetic and clinical baseline variables. We

compare the predictive ability of the algorithms used and evaluate the strengths and weaknesses of their application to pharmacogenetic problems.

# 1. Introduction

## 1.1. Overview of pharmacogenetics

When it comes to pharmacotherapy, one size does not fit all. Whereas some patients gain improvements in their condition as intended, others may fail to respond clinically or suffer from an adverse drug reaction (ADR), which results in treatment failure (Fig. 1). Non-response to drug treatment and the occurrence of ADRs is a major burden on patients and society as a whole. On average, for 25-50% of patients, drug treatment response is suboptimal across a wide range of treatments (Spear, Heath-Chiozzi, & Huff, 2001). Furthermore, it is estimated that 3.5% of hospital admissions in Europe are caused by an ADR and that 10% of hospitalized patients develop and ADR (Bouvy, De Bruin, & Koopmanschap, 2015). On top of the health risks imposed on patients, ADRs have considerable economic consequences. On average ADRs increase the cost of hospitalization by US$2,000 or 20%, and the yearly cost of ADRs may exceed US$30 billion in the US and €24 billion in Europe (European Commission, 2008; Khan, 2013; Sultana, Cutroneo, & Trifirò, 2013). The duration of hospitalization increases by 8% in patients suffering an ADR (Khan, 2013). However, up to 20-30% of ADRs could be prevented by pharmacogenetic (PGx) testing (Alfirevic & Pirmohamed, 2017). Clearly, a predictive biomarker that personalizes treatment and improves response rates or prevents a proportion of ADRs from occurring would save health care resources and greatly improve patient experience.

The variability in how patients respond to drugs is partly due to differences in patients' characteristics such as age, sex, ethnicity and genetic make-up, and environmental factors such as diet, medication and alcohol and tobacco use. The research field of PGx studies the association between genetic markers and the variability in response to drug treatment. The terms pharmacogenetic and pharmacogenomic imply the study of interactions between

drug response and a single gene or the whole genome, respectively, but are used interchangeably in the literature and in this thesis (PharmGKB, 2017a). Genetic biomarkers have been associated with efficacy of drug treatment, risk of developing ADRs and drug dosage, though these distinctions are not always clear cut as toxicity may be the result of excessively high dosing and non-response due to inadequately low dosing. Different types of genetic variation exist, for example single nucleotide polymorphisms (SNPs), deletions, insertions or short tandem repeats, though in this PhD we mainly focus on SNPs.

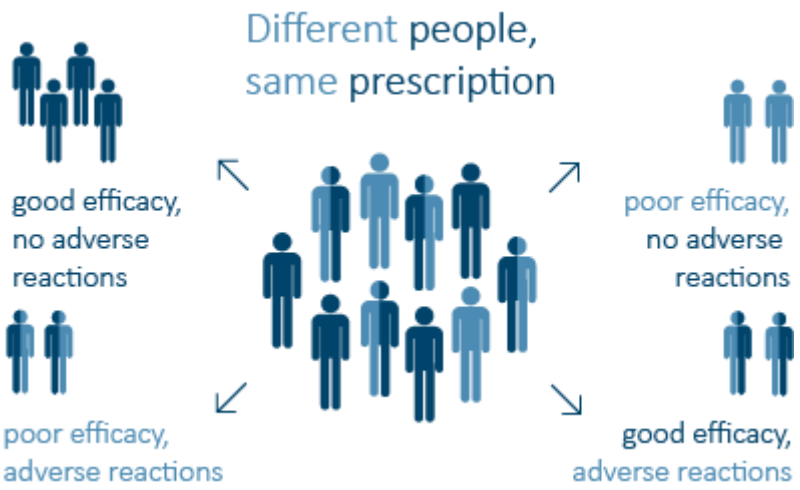



Figure 1. Inter-individual variability in drug response. Adapted from Samwald (2017).

PGx variants are used as predictive biomarkers since they predict how a patient will respond to treatment, in contrast to prognostic biomarkers which predict disease progression independently of treatment. A patient's genetic profile, alone or in combination with demographic details, clinical observations and environmental factors, can thus sometimes be used to tailor treatment to that individual patient. The hope is that personalized medicine will reduce the number of adverse events, and improve and quicken treatments by selecting the most effective drug and choosing the optimal dose (Jorgensen & Pirmohamed, 2011). In addition to improving patients' health outcomes, PGx discoveries may increase our understanding of the pharmacodynamics and pharmacokinetics of a drug.

An important advantage of using DNA to personalise treatment is that DNA sequences largely remain unaltered over time, unlike gene expression or methylation levels and metabolic biomarkers. By consequence, if the genetic code of a patient is obtained once, this information stays relevant throughout their life. It is conceivable that whole genome sequencing, which currently costs approximately £1,000, would become cheap enough to be routinely incorporated in health records (Wetterstrand, 2016). Genotyping arrays, covering SNPs spread across the genome, are less costly (approximately £30-£200) and can also contain useful PGx information (Lu, Lewis, & Traylor, 2017). An exception to the unchangeable nature of DNA is somatic mutations in tumour cells. The DNA sequence of cancerous cells can mutate rapidly, so this can be regarded as a separate branch of PGx which will not be investigated in this thesis. However, there has been a lot of progress in this field and many cancer drugs are accompanied by PGx tests (Patel, 2016).

The search for new predictive biomarkers often starts by investigating candidate genes, typically genes involved in the pharmacokinetic and pharmacodynamic pathways of a drug (Fig. 2). Since PGx effects of a variant only become evident after exposure to a drug, which may not happen for many carriers, these genes are not under selection pressure (Roden, Wilke, Kroemer, & Stein, 2011). Hence, PGx variants can be relatively common in the population. For example, the cytochrome P450 (*CYP450*) enzymes *CYP1A2*, *CYP2D6*, *CYP2C9*, *CYP2C19* and *CYP3A4* are involved in the metabolism of 60% of drugs and variants of these enzymes with increased or decreased enzyme activity are common in the Caucasian population (Table 1) (Preissner et al., 2013). A study looking at five commonly used drugs found that 91% of patients in the United States carries at least one variant that affects drug response (Van Driest et al., 2014). The drugs (and respective genes) included in this study were clopidogrel (*CYP2C19*), simvastatin (*SLCO1B1*), warfarin (*CYP2C9* and *VKORC1*), thiopurines (*TPMT*), and tacrolimus (*CYP3A5*).

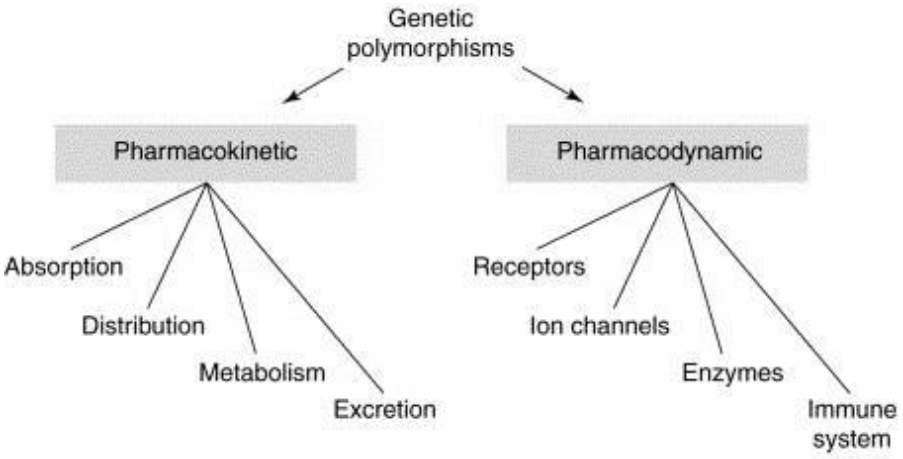



Figure 2. Pharmacokinetic and pharmacodynamic pathways can be affected by genetic variability. Adapted from Pirmohamed & Park (2001).

Table 1. Frequency of selected *CYP* variants in the Caucasian population. Adapted from Preissner et al. (2013) .

| ***CYP*** | **Allele** | **Frequency (%)** | **Enzyme Activity** |
|---|---|---|---|
| *1A2* | **1F* | 33.3 | Higher inducibility |
| | **1D* | 4.82 | Decrease |
| *2C9* | **2* | 19.0% **1/*2*; 1.6% **2/*2*; 1.8 % **2/*3* | Decrease |
| | **3* | 9 | Decrease |
| *2C19* | **2* | 16 | Decrease |
| | **17* | 18 | Increase |
| *2D6* | **3* | 2.04 | Decrease |
| | **4* | 20.7 | Decrease |
| | **4D* | 3.4 | Decrease |
| | **4L* | 4.5 | Decrease |
| | **5* | 4.1 | no enzyme |
| | **6* | 1.3 | Non-functional |
| | **7* | 1 | Decrease |
| | **9* | 2 | Decrease |
| | **10* | 8 | Decrease |
| | **41* | 8 | decrease (expression) |
| *3A4* | **17* | 2 | Decrease |
| | **1B* | 17 | increase (transcription) |
| | **2* | 2.7 | Decrease |

It is important to note that not all statistically significant PGx biomarkers can be translated to clinically useful predictive tests. A PGx variant indicating a small increase in risk of developing an ADR or slightly lower chance of response might not have a clinical impact on the choice of drug. The required effect size to be of clinical relevance depends on specific circumstances, such as the severity of non-response or the ADR, the absolute risk of non-response or ADR development and the availability and efficacy of alternative drugs. In order to have clinical utility, the effect size of a PGx variant must provide relevant and actionable information to the physician.

In addition to clinical utility, the economic impact of PGx testing needs to be evaluated before implementing a biomarker in clinical practice. As health care resources are limited, the economic consequences of applying new technologies are important. Health economic studies compare costs and outcomes of different treatment strategies to assess which option is more cost-effective. A treatment is deemed cost-effective if the gain in health outcomes reduces total costs or comes at an affordable increase in costs. A review of the economics of individualized medicine found that in general personalized medicine is as cost-effective as other health care interventions (Hatz, Schremser, & Rogowski, 2014). This study included screening and prognostic biomarkers in addition to predictive genetic tests and also covered tests of viral or bacterial DNA. However, predictive PGx testing was more expensive than the use of prognostic or screening biomarkers. A second literature review of personalized medicine found that the majority of genetic tests were cost-effective and one in five tests reduced costs while providing better health outcomes (Phillips et al., 2014). The economics of PGx testing is a rapidly evolving field as more genetic biomarkers are discovered and the cost of genetic testing decreases.

### 1.1.1. Pharmacogenetic information and drug labelling

PGx information that is deemed relevant is sometimes included on the drug label or package insert. The US Food and Drug Administration (FDA) curated Table of Pharmacogenomic Biomarkers in Drug Labeling lists 204 biomarkers on the labels of 165 unique drugs in various treatment areas (U.S. Food and Drug Administration, 2017b). Information on the biomarker is included in the drug label, though not all labels provide guidance regarding genetic testing. One third of drugs on the FDA table are used in oncology (52 drugs, 32%), while psychiatry and infectious diseases are the second and third largest therapy areas with respectively 16% (27 drugs) and 10% (16 drugs) of PGx biomarkers (Fig. 3). *CYP450* enzymes are involved in the metabolism of many drugs and are included frequently on FDA drug labels (Fig. 4). *CYP2D6* has PGx associations with 44 drugs, while *CYP2C19* is mentioned in 19 and *CYP2C9* in 6 drug labels. Although *CYP2D6* is the most quoted genetic biomarker on FDA drug labels, it has not had a significant impact on clinical practice (Pirmohamed, 2014). The glucose-6-phosphate dehydrogenase (*G6PD*) gene is associated with 22 drugs, among which antimalarials, analgesics and antibiotics (Luzzatto & Seneca, 2014). With regards to psychiatry, it is remarkable that all 27 psychiatric drugs listed on the FDA table have PGx associations with *CYP2D6*, and 3 drugs also with *CYP2C19*.

A second resource, the PharmGKB website managed by Stanford University, identifies 202 drugs that contain PGx information on the FDA drug label and 92 with PGx details on the European Medicines Agency (EMA) European Public Assessment Report (Whirl-Carrillo et al., 2012). Genetic testing prior to use of the drug is required for 50 and 35 drugs by the FDA and EMA, respectively (Fig. 5). For a further five (FDA) and two (EMA) drugs genetic testing is recommended on the label. Differences in regulation and labelling between the US FDA and the European EMA could partly be due to varying allele frequencies in the population, which impact sensitivity and specificity of the PGx biomarker.

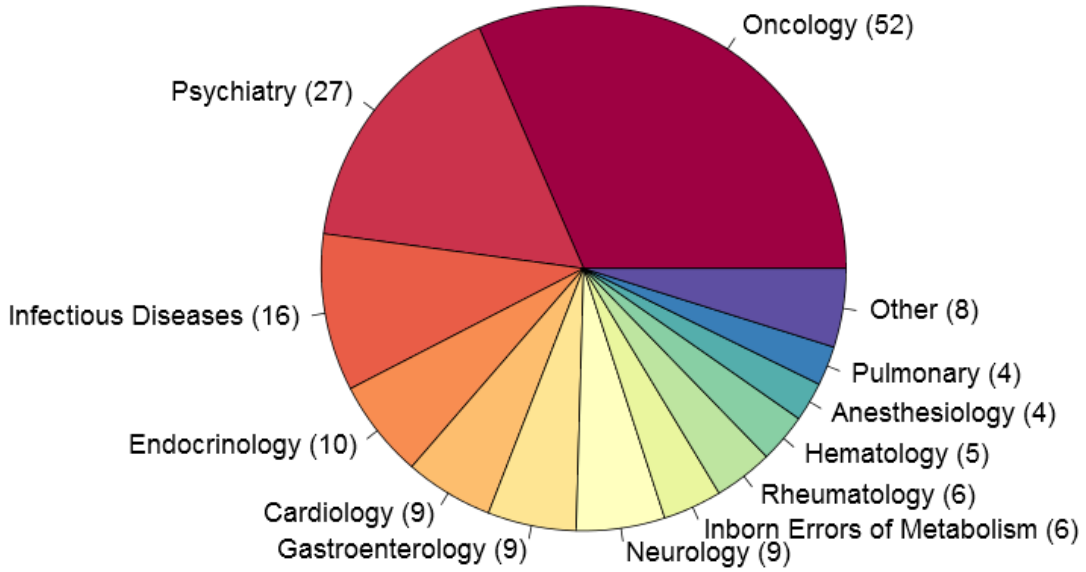



Figure 3. Therapeutic areas of drugs listed on FDA Table of Pharmacogenomic Biomarkers in Drug Labeling (U.S. Food and Drug Administration, 2017b).

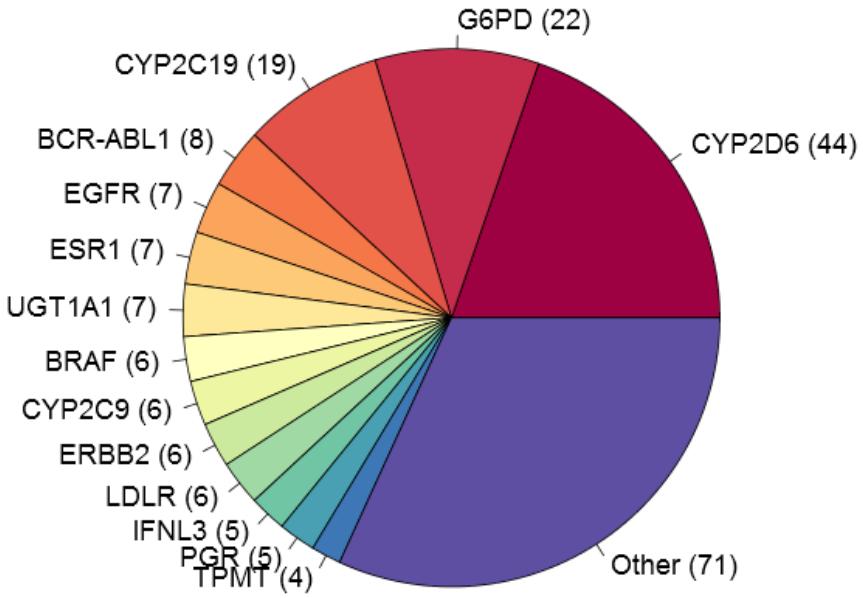



Figure 4. Genes reported in labels of drugs listed on FDA Table of Pharmacogenomic Biomarkers in Drug Labeling (U.S. Food and Drug Administration, 2017b).

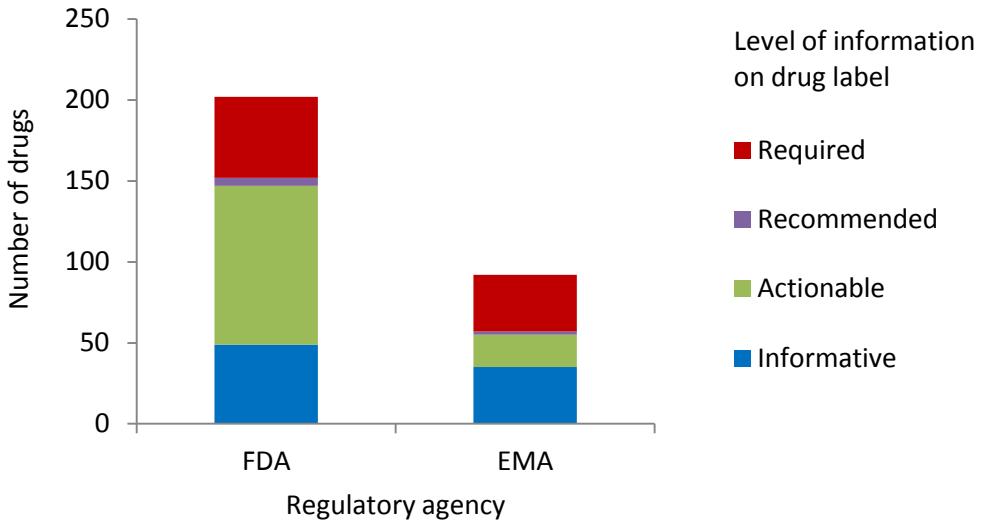



Figure 5. Requirements on drug labels regarding PGx testing based on information from PharmGKB.org (PharmGKB, 2017b; Whirl-Carrillo et al., 2012).

### 1.1.2. Examples of pharmacogenetic biomarkers

There are some well-established PGx associations which are used in clinical practice as predictive PGx tests to improve drug safety and efficacy. An example of a safety PGx biomarker is the association of the HIV-1 reverse transcriptase inhibitor abacavir with a human leucocyte antigen (*HLA*) variant. About 4% of patients treated with abacavir develop a potentially fatal hypersensitivity syndrome (Hetherington et al., 2002; Symonds et al., 2002). *HLA-B*5701* is strongly associated with abacavir hypersensitivity syndrome with an exceptionally large odds ratio of 117 (Mallal et al., 2002). The clinical usefulness of *HLA-B*5701* screening has been confirmed in a clinical trial and as a result *HLA-B*5701* screening before initiation of abacavir treatment is required in the UK and US (Electronic Medicines Compendium, 2016b; Hughes, Hughes, Brothers, Spreen, & Thorborn, 2008; Mallal et al., 2008; U.S. Food and Drug Administration, 2015).

A second successful example of PGx safety associations with *HLA* alleles is the anti-epileptic drug carbamazepine. Carriers of the *HLA-B*1502* allele are at high risk of developing Stevens-Johnson syndrome and toxic epidermal necrolysis, a severe ADR (Lim, Kwan, & Tan, 2008). This association is very strong in Han Chinese and Taiwanese patients, with odds ratios exceeding 1000, and was also detected in Malay and Thai samples (Chung et al., 2004; Hung et al., 2006). The *HLA-B*1502* allele has low frequency in Japanese, Korean, European and African populations and occurrence of the ADR is rare in these patients (Lim et al., 2008). However, in Caucasian patients, *HLA-A*3101* has been linked to Stevens-Johnson syndrome and toxic epidermal necrolysis (McCormack et al., 2011). PGx testing is mandatory for patients of Asian descent prior to starting carbamazepine treatment in the US and testing is strongly recommended in the UK (Electronic Medicines Compendium, 2015; U.S. Food and Drug Administration, 2013b) . The drug label also mentions the *HLA-A*3101* association without testing requirements.

In addition to reducing the risk of ADRs, genetic variants may be used as efficacy biomarkers to identify patients who are likely to respond to a drug. Ivacaftor is used for the treatment of cystic fibrosis, a disease caused by over 1900 different mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) protein (Lubamba, Dhooghe, Noel, & Leal, 2012). The efficacy of ivacaftor depends on specific mutations in the *CFTR* gene. Cystic fibrosis patients with at least one *G551D-CFTR* allele benefit from ivacaftor, but the drug is ineffective for patients who are homozygous for *the F508del-CFTR* variant (Davies et al., 2013; Flume et al., 2012; Ramsey et al., 2011). Both the FDA and the EMA require genetic testing before starting ivacaftor treatment (Electronic Medicines Compendium, 2016a; U.S. Food and Drug Administration, 2017a).

The anti-coagulant warfarin is an example of how a PGx test can be used to establish the appropriate dose of a drug. The therapeutic dosing of warfarin is closely monitored by measuring the standardized international normalized ratio (INR), but there is wide variation

between patients in the dose needed to achieve a therapeutic INR. Variants of the vitamin K epoxide reductase complex (*VKORC1*), the target to which warfarin binds, and *CYP2C9* and *CYP4F2*, involved in the metabolism of the drug, have been associated with therapeutic warfarin dose. Together these genes explain around 40% to 54% of variability between patients (Fisch, Perry, Stephens, Horenstein, & Shuldiner, 2013). However, four clinical trials comparing time in therapeutic INR range between patients who were prescribed warfarin based on a dosing algorithm including *CYP2C9* and *VKORC1* genotypes and clinical variables, versus an algorithm based on clinical variables only did not find significant differences (Jonas et al., 2013; Kimmel et al., 2013; Pengo et al., 2015; Verhoef et al., 2013). A fifth trial concluded that the PGx algorithm lead to a higher percentage of time in therapeutic INR range and also that this range was reached more quickly (Pirmohamed et al., 2013). Although there is still uncertainty about the clinical utility of PGx guided warfarin dosing, the FDA recommends including *CYP2C9* and *VKORC1* genotype information in the calculation of the starting dose, if this information is available (U.S. Food and Drug Administration, 2016a).

## 1.2. Research methods in pharmacogenetics

The scientific methods used to investigate PGx associations have made a lot of progress, both in terms of the range of genetic variants studied as well as statistical rigour. The earliest PGx studies used the ratio between a drug and its metabolites to measure enzyme activity. This allowed distinguishing between genetic variants that cause different enzyme activity phenotypes. The assay method was for example used in the discovery of *CYP2D6* poor and ultrarapid metabolizers (Ingelman-Sundberg, 2005). In fact, phenotyping is still used in clinical practice to determine *CYP2D6*, *TPMT* and *G6PD* variants (Pirmohamed, 2014).

The development of the polymerase chain reaction (PCR) technique and sequencing made it possible to study specific candidate genes. The discovery of the *HLA-B*5701* allele with abacavir hypersensitivity is a successful example of this approach. However, few findings from candidate gene studies have been replicated and many associations are likely false positives. This is partly due to the small sample sizes typically used in these PGx studies. In disease genetics studies increasing the sample size often revealed that effects estimated in small samples were overly optimistic (Fig. 6). In addition, candidate gene studies focused on one or a few variants in a gene instead of the variability in the entire gene (Pirmohamed, 2014). Furthermore, the mechanism of action of a drug might not be fully clear and this complicates selection of candidate genes.

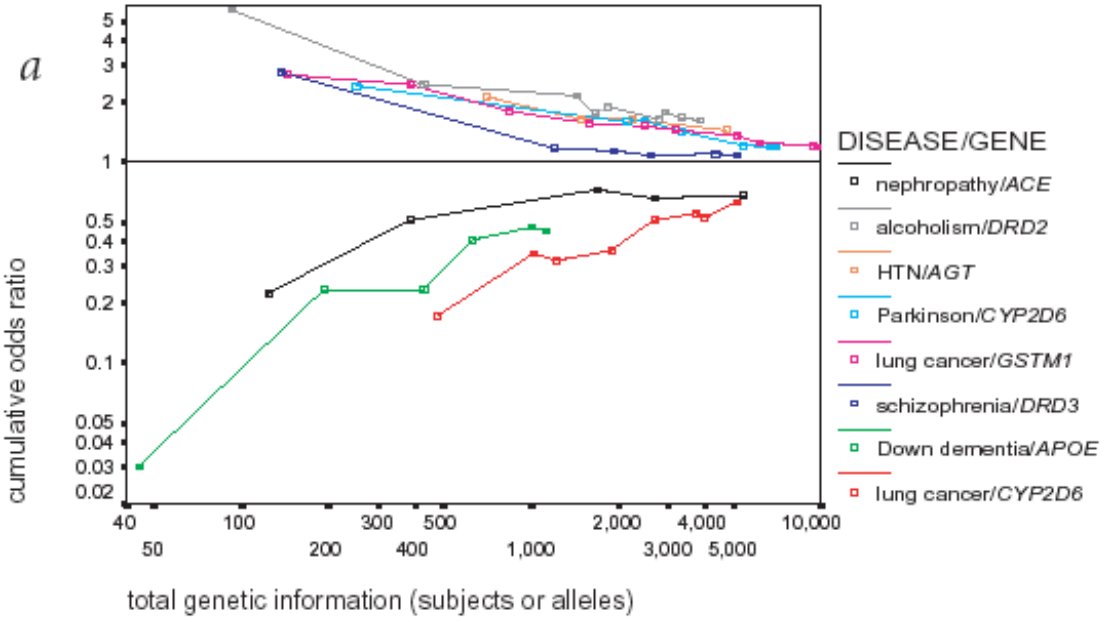



Figure 6. Strength of association decreases with increasing sample sizes. Adapted from Ioannidis, Ntzani, Trikalinos, & Contopoulos-Ioannidis (2001).

The question of choosing genes to study does not present itself in genome-wide association studies (GWAS), as they take a hypothesis free approach. Millions of variants can be assessed and hence *p*-values are compared to a strict statistical significance threshold

(usually $10^{-8}$) to account for multiple testing. By consequence, sample sizes need to be large to have sufficient statistical power to detect associations, particularly for moderate effect sizes. This partly explains why the success of PGx GWAS has been modest so far. GWAS of PGx outcomes have typically collected hundreds to thousands of patients, whereas some disease GWAS samples include over 100,000 individuals. In disease GWAS, increased sample sizes have led to more significant findings and the same is expected in PGx research (Fig. 7). However, it can prove difficult to recruit participants for PGx studies who not only need to suffer from the same disease, but also need to be treated with the same drug. In addition, studying a rare ADR can make it more challenging to attain a substantial sample size. International and multi-centre collaborations help to increase sample sizes. The field of PGx GWAS is not as advanced as disease genetics and fewer GWAS have been conducted (Fig 8). More and larger PGx GWAS are needed to confirm the associations from candidate gene studies and to uncover new PGx markers. Moreover, next generation sequencing techniques such as whole exome and whole genome sequencing allow even denser coverage of genetic variants (Katsila & Patrinos, 2015). This opens the door to studying a greatly increased number of variants, including rare variants.

## 1.3. Pharmacogenetics in psychiatry

Although various PGx associations with psychiatric drugs have been discovered, few associations have effects strong enough to influence clinical practice (Arranz & Kapur, 2008). Many psychiatric drugs are metabolized by *CYP450* enzymes and may inhibit or induce enzyme activity. *CYP2D6* and *CYP2C19* are the genes with the most clinically relevant PGx associations (Table 2) (Spina & de Leon, 2015). *CYP2D6* variants are for example used in dosing guidelines for venlafaxine, pimozide, aripiprazole, haloperidol and tricyclic antidepressants. Patients treated with citalopram or escitalopram who are ultrarapid *CYP2C19* metabolizers should receive increased doses to achieve the same level

of efficacy (Swen et al., 2011). Furthermore, response to antipsychotic treatment is associated with dopamine receptor (*DRD2* and *DRD3*), serotonin receptor (*HTR1A* and *HTR2A*) and zinc-finger domain-containing protein (*ZNF804A*) genes (Pouget, Shams, Tiwari, & Müller, 2014). *DRD2* and *HTR2A* have also been found to be associated with tardive dyskinesia, an ADR to antipsychotics. However, none of these genetic variants provides clinically relevant predictions on their own as effect sizes are small. The sample sizes collected in candidate gene studies have typically been small, so there is a considerable risk of false positive findings. Therefore, these PGx variants should be replicated, for example in candidate gene studies or GWAS, to provide more robust evidence for association.

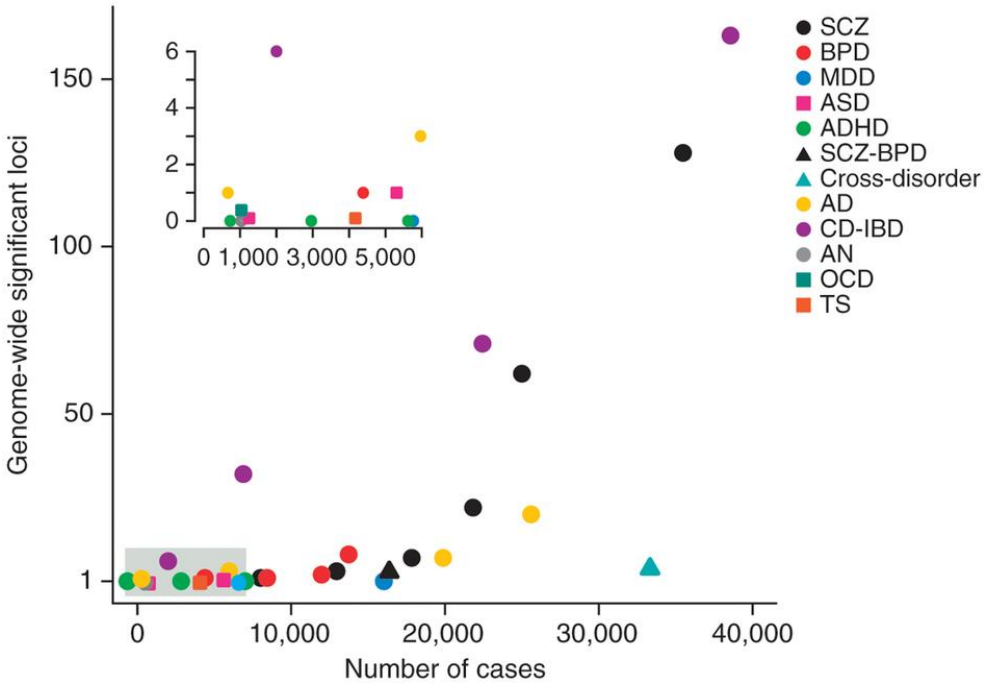



Figure 7. Number of independent genome-wide significant loci as a function of the numbers of cases in the largest meta-analysis. Cross-disorder indicates a broad psychiatric phenotype. AD: Alzheimer's disease; ADHD: attention deficit hyperactivity disorder; AN: anorexia nervosa; ASD: autism spectrum disorder; BPD: bipolar disorder; CD-IBD: Crohn's disease and inflammatory bowel disease; MDD: major depressive disorder; OCD: obsessive compulsive disorder; SCZ: schizophrenia; TS: Tourette's syndrome. Adapted from Gratten, Wray, Keller, & Visscher (2014).

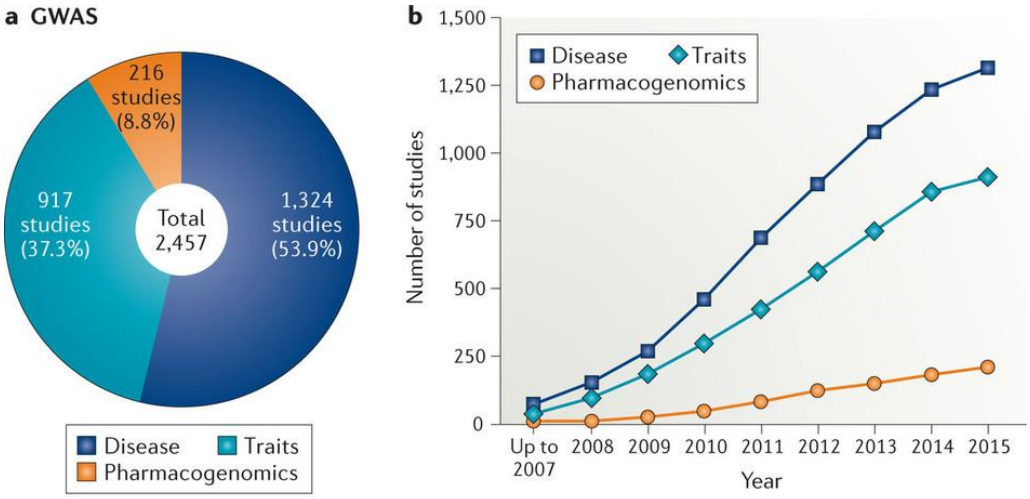



Figure 8. Number of GWAS (a) in total and (b) per year in human diseases, traits and PGx. Adapted from Giacomini et al. (2017).

Table 2. Dosing guidelines for psychiatric drugs using *CYP2C19* and *CYP2D6*. Adapted from Spina & de Leon (2015).

| **Drug** | **Dosing guidelines** |
|---|---|
| Aripiprazole, haloperidol, risperidone and zuclopenthixol | *CYP2D6* PMs: ↓ dose by 50 % or select another antipsychotic |
| | *CYP2D6* UMs: be alert to diminished serum concentrations or prescribe another antipsychotic |
| Atomoxetine | *CYP2D6* UMs: be alert to reduced efficacy or select alternative drugs |
| Long-acting intramuscular aripiprazole | *CYP2D6* PMs: ↓ dose to 75 % |
| Pimozide | If prescribing >4 mg/day in adults, *CYP2D6* genotyping is required by the US prescribing information because 4 mg/day is the maximum recommended dose in *CYP2D6* PMs |
| Tricyclic antidepressants | *CYP2D6* PMs: avoid TCAs or ↓ dose by 50 % and use TDM to adjust dosing |
| | *CYP2D6* UMs: avoid TCAs |
| | *CYP2C19* PMs and amitriptyline: ↓ dose by 50 % and use TDM to adjust dosing |
| | *CYP2C19* UMs and amitriptyline: select another antidepressant not metabolized by *CYP2C19* |
| Venlafaxine | *CYP2D6* PMs: select another antidepressant or use venlafaxine TDM |
| | *CYP2D6* UMs: increase dose by a factor of 1.5 |

PM: poor metabolizer; TCA: tricyclic antidepressant; TDM: therapeutic drug monitoring; UM: ultrarapid metabolizer

Most PGx studies in psychiatry, as in other therapy areas, were candidate gene studies, but recently GWAS have been used to expand the search for genetic variants. A GWAS of antipsychotic induced weight gain found a significantly associated SNP in the melanocortin 4 receptor (*MC4R*) that was confirmed in 3 replication cohorts (Malhotra, Correll, et al., 2012). Although the first GWAS studying lithium response in bipolar patients (STEP-BD study) did not identify any genetic associations (Perlis et al., 2009), recent studies were more successful. Four linked SNPs in a region containing long, non-coding RNAs on chromosome 21 were significantly associated with lithium response in a European and Asian sample (Hou et al., 2016). A GWAS in Swedish and UK samples compared lithium responsive patients with healthy controls and detected a SNP in the 6-pyruvoyltetrahydropterin synthase gene (*PTS*), yet no SNPs were significantly associated with response among lithium treated patients (Song et al., 2017; Song et al., 2016). In addition, a two SNP signal in the glutamate decarboxylase-like protein 1 gene (*GADL1*) was associated with lithium response with an extremely significant $p$-value ($p$=5.50×10$^{-37}$) and could potentially be a highly accurate predictor of treatment response (Chen et al., 2014). However, this association has only been confirmed in Han Chinese sample but not in other populations (Pickard, 2017). Five GWAS of antidepressant response, the GENDEP, MARS, STAR*D, PGRN-AMPS and ISPC studies, did not find any significant associations (Biernacka et al., 2015; Garriock et al., 2010; GENDEP Investigators, MARS Investigators, & STAR*D Investigators, 2013; Ising et al., 2009; Ji et al., 2013; Uher et al., 2010). Polygenic risk scores also did not predict treatment response from the GENDEP to STAR*D study or vice versa (García-González et al., 2017). Furthermore, GWAS studying antipsychotic response (CATIE studies) did not detect statistically significant variants (Malhotra, Zhang, & Lencz, 2012). The small sample sizes used in these GWAS contribute to the lack of success in identifying PGx variants, though it is expected that larger studies will lead to more genome-wide significant findings (Giacomini et al., 2017).

Few PGx associations are translated to widely used predictive tests, because large effect sizes are needed to achieve clinical utility. A PGx test based on multiple *CYP450* genes, serotonin transporter (*SLC6A4*) and serotonin receptor (*5-HTR2A*) genes (GeneSight test by Assurex Health) to guide antidepressant treatment has been developed commercially and claims to improve treatment response compared to treatment as usual (Altar et al., 2015; Hall-Flavin et al., 2012; Hall-Flavin et al., 2013; Winner, Carhart, Altar, Allen, & Dechairo, 2013). Another test using *CYP450*, *UGT1A1* and *ABC* transporter genes to guide antidepressant dosing (CNSDose test by CNSDose) improved remission compared to standard treatment (Bousman et al., 2017; Singh, 2015). Yet the utility of these tests should be judged with some caution, as the studies supporting the efficacy claims had small sample sizes ($n$ = 227, 44 and 51 for the GeneSight test and $n$ = 148 and 119 for the CNSDose test) and have all been conducted by the companies marketing the tests. The AmpliChip *CYP450* test (Roche Molecular Systems) identifying *CYP2D6* and *CYP2C19* variants was the first FDA approved PGx test and can be used to guide antipsychotic drug selection (Pouget et al., 2014). However, tests for clozapine response (using the *5-HT2A* and *5-HT2C* serotonin receptor, *H2* histamine receptor, and the *5-HTT* serotonin transporter genes) and clozapine-induced agranulocytosis (using a SNP in *HLA*DQB1*) have been marketed but are no longer available (Kohlrausch, 2013; Pouget et al., 2014). The moderate effect sizes of these tests did not lead to changes in clozapine treatment.

## 1.4. Machine learning for predictive modelling

Genetic studies collect large datasets spanning thousands or millions of variants, many of which are correlated. The number of study participants is often considerably smaller than the number of variables. These data characteristics complicate the statistical analysis of genetic studies. Typically, statistical methods are applied to individual genetic variants one by one and significance thresholds are corrected for multiple testing. However, these

univariate approaches do not take interactions between the variables into account. Due to the stringent significance thresholds applied genetic variants with weak effects can easily be missed. However, treatment response and ADRs could well be caused by a combination of multiple variants with moderate or weak effect sizes and it is possible that up to $10^4$ genetic variants contribute to these traits (Pouget et al., 2014). Therefore, it is appropriate to perform a multivariable analysis and assess several genetic variants simultaneously.

Machine learning is a discipline integrating concepts from statistics and computer science to build flexible multivariable prediction models. These methods enable the analysis of large datasets such as genome-wide studies with millions of genetic variants in a single model. In addition, machine learning algorithms optimize prediction accuracy and are thus well suited for the development of predictive biomarkers. Some machine learning methods, for example elastic net, inherently perform variable selection and can help identifying relevant genetic variants from a genome-wide dataset.

Several machine learning methods, including elastic net, random forest, support vector machines and neural networks, have been used to build prediction models for warfarin dosing in different populations. Machine learning algorithms proved successful in predicting optimal warfarin dosing based on clinical characteristics or a combination of clinical and genetic variables (Cosgun, Limdi, & Duarte, 2011; Grossi et al., 2014; Liu, Li, Zhang, & Zhou, 2015; Pavani et al., 2016; Sharabiani, Bress, Douzali, & Darabi, 2015). However, other studies found that multiple linear regression outperformed machine learning algorithms (International Warfarin Pharmacogenetics Consortium, 2009; X. Li et al., 2015). Machine learning was also successfully applied to construct different prediction algorithms for tacrolimus dose in renal transplant patients (Tang et al., 2017). Furthermore, tree-based machine learning algorithms were able to identify patients most likely to benefit from pharmacotherapy in an alcohol dependence trial (Hou et al., 2015). Although machine learning methods can perform well with large datasets, the PGx studies mentioned above

included only a small set of previously associated genetic variants as predictors. A study combining clinical and genetic predictors used elastic net to predict antidepressant response to a clinically relevant degree (Iniesta et al., 2015). This study did include genome-wide variants and allowed the algorithm to select the most relevant features for inclusion in the prediction model.

In the last few years, computer processing power has increased rapidly. Combined with the availability of ever larger datasets, this has fostered advances in deep learning, a branch of machine learning covering neural networks and their variations. Deep learning applications are the algorithms underlying artificial intelligence and are widely used. For example, online recommendation systems, which apply algorithms to web traffic data and suggest items based on what previous users have looked at, demonstrate how deep learning is routinely used in everyday life (Cheng et al., 2016; Covington, Adams, & Sargin, 2016). A second well-known example is image recognition, where deep learning algorithms can recognise and identify faces nearly as well as humans can (Parkhi, Vedaldi, & Zisserman, 2015; Taigman, Yang, Ranzato, & Wolf, 2014). Other areas where deep learning significantly progressed the field are speech recognition and software for self-driving cars (Deng et al., 2013; Reiley, 2016). Medicine has been named as the next frontier in deep learning (Frey, 2016). Automated image classification tools can assist in diagnosis making, for example in oncology where neural networks could be used to detect prostate and breast cancer in histopathological images of biopsies (Litjens et al., 2016). Furthermore, a deep learning network outperformed dermatologists in classifying cancers from skin lesion images and the authors suggest that their algorithm could be made available on smartphones as a low-cost diagnostic tool (Esteva et al., 2017).

Machine learning and deep learning approaches provide novel tools for the analysis of large genetic datasets. A strong advantage of these methods over traditional statistical models is that they enable studying multiple correlated genetic variants simultaneously. Moreover, as

machine learning algorithms aim to optimize prediction accuracy, they can readily be applied as tests for personalizing treatments.

## 1.5. Outline of thesis

The research presented in this thesis is grouped in two parts. Firstly, we look into aspects related to the translation of PGx associations to genetic tests that are implemented in clinical practice to guide pharmaceutical treatment. In the second part, we apply machine learning techniques to genetic and gene expression datasets to predict clinical outcomes.

### 1.5.1. Adoption of pharmacogenetic testing in clinical practice

The adoption of novel PGx biomarkers in clinical practice is not self-evident. We studied two factors that are relevant to the transition from bench to bedside: the predictive power that a PGx test needs to achieve in order to be clinically useful and the economic feasibility of PGx testing. The research introduced in the following paragraphs led to two peer-reviewed journal publications and an invited editorial. The research chapters in this first part of the thesis consist of these manuscripts.

Firstly, we studied the characteristics of a PGx test for clozapine induced agranulocytosis, a rare but severe ADR. Clozapine in an effective antipsychotic, but due to the risk of agranulocytosis the drug is reserved for treatment resistant schizophrenia and patients must undergo regular haematological monitoring throughout the course of treatment. A PGx test that stratifies patients into agranulocytosis risk subgroups could have a big impact on schizophrenia pharmacotherapy. We examined the relationship between test sensitivity and the proportion of positive test results, which is related to the allele frequency of the genetic variant(s) (Verbelen, Collier, Cohen, MacCabe, & Lewis, 2015). In light of our findings, we revisited the previously established PGx associations with clozapine induced agranulocytosis and comment on the lack of clinical impact they made (Verbelen & Lewis,

2015). The framework we developed to assess the utility of PGx testing for clozapine induced agranulocytosis can readily be adapted to other settings where PGx biomarkers for drug response or safety are evaluated.

Secondly, we looked at the economic arguments for PGx testing. In addition to clinical utility, economic criteria need to be fulfilled to warrant the use of a PGx test as standard practice. Healthcare budgets are limited, thus the cost of an intervention needs to be balanced with its benefits. Although the use of a PGx test is intended to decrease the rate of non-response or ADRs, this does not automatically reduce costs. In addition to the price of the genetic test itself, the costs and health outcomes of the treatment following the test result need to be considered, for example the cost and efficacy of an alternative drug. Economic evaluations or cost-effectiveness studies assess and compare the costs and benefits of different healthcare strategies. We reviewed the health-economic literature to form a picture of the cost-effectiveness of PGx tests (Verbelen, Weale, & Lewis, 2017). As the cost of genotyping keeps decreasing, it is conceivable that genetic information might become part of patients' health records and could be used for PGx guidance without additional cost. Therefore, we estimated the impact of freely available genetic information on the cost-effectiveness conclusions of the reviewed publications.

### 1.5.2. Machine learning prediction algorithms applied to genetic and gene expression studies

The second part of this thesis emphasises the application of machine learning to study genetic and gene expression data. Due to the size of genetic datasets, genetic variants are traditionally analysed on a one by one basis. In contrast, machine learning algorithms can be fit with large numbers of variables and can thus be used to perform multivariable analyses where all genetic variants are studied in a single model. As the focus of machine learning models lies on prediction, these methods are ideal to build polygenic prediction

models, potentially including clinical variables. In addition, variable importance scores can be used to rank predictor variables and derive which genetic variants are most predictive of the outcome.

We used machine learning methods to perform multivariable analyses of genetic and gene expression studies. Various statistical approaches were applied and we start by giving an overview of the statistical methods used. Next, we describe three studies that were undertaken using machine learning algorithms.

In the first study, machine learning algorithms were built to classify schizophrenia cases from controls using RNA sequencing (RNA-seq) gene expression data from post-mortem dorsolateral prefrontal cortex (DLPFC) brain samples. The dataset was randomly split in training data for model building and test data for assessing the predictive accuracy of the algorithm. Random forests, support vector machines and neural networks were trained to distinguish cases from controls. The predictive performance of the algorithms was assessed on the test data. Using different feature selection methods, we investigated whether a smaller set of genes can achieve comparable or higher prediction accuracy than an algorithm based on the entire set of predictor variables. Thus, we tested if machine learning algorithms can be used to identify sets of genes of which the expression levels contribute most to predicting case/control status.

Secondly, we applied linear and tree-based machine learning algorithms to a phase II anti-diabetic clinical trial. This was a cross-over trial comparing a glucagon receptor antagonist (LY2409021) to placebo in type 2 diabetes patients. Our goal was to construct predictive PGx algorithms for five continuous anti-diabetic efficacy and safety measures using clinical variables and genome-wide SNPs. Again, the available data were randomly split in training and test subsets. Regression trees, random forest and elastic net algorithms were built using the training data and their predictive performance was assessed in the independent

test data. Furthermore, feature selection techniques were used to improve the signal to noise ratio in the predictor variables, which may enhance the performance of the algorithms. A sensitivity analysis was carried out to verify the robustness of the results against random sampling errors in to the training/test data splitting process.

Thirdly, we contrasted traditional statistical analysis methods with machine learning algorithms for the analysis of a phase II clinical trial comparing a highly selective norepinephrine reuptake inhibitor (LY2216684) to placebo in patients suffering from major depressive disorder. We used clinical and genetic variables in candidate genes to model changes in the Montgomery-Åsberg depression rating scale total score (MADRS-TS), a measure of depression severity. The aim of this study was to identify SNPs that have an effect on changes in the outcome variable. First, we looked at cross-sectional methods, namely linear regression and elastic net, to model the change from baseline in the outcome variable at the end of the trial period. Next, we approached this analysis from a longitudinal perspective and used linear mixed models and linear mixed elastic net to model changes in MADRS-TS over the course of the trial. We compared the results from the traditional and machine learning analysis methods. In addition, we reviewed the existing software packages for longitudinal machine learning.

### 1.5.3. Summary

This thesis studies how genetics can be used to improve pharmacological treatment across a range of disorders. In the research performed, PGx biomarkers are approached from two different angles. Firstly, we assess the adoption of PGx associations as predictive tests in clinical practice. We examined the characteristics of a clinically useful PGx test for clozapine induced agranulocytosis and reviewed the literature of cost-effectiveness studies of PGx tests. Secondly, we applied machine learning algorithms to genetic and gene expression datasets to evaluate the utility of these methods for building algorithms that accurately

predict clinical outcomes. We also investigated how these methods can be used to identify risk variants that would play a role in response to treatment or in adverse events.

# Part 1: Adoption of pharmacogenetic testing in clinical practice

The first part of this thesis investigates aspects related to the clinical use of PGx tests. These chapters consist of two published peer-reviewed articles and an editorial. Firstly, we examined the characteristics of a PGx test for clozapine induced agranulocytosis. In the accompanying editorial we comment on why the known PGx associations are not yet used to guide clinical practice and highlight the challenges in this research area. Secondly, we performed a review of the pharmaco-economic literature on PGx testing to get a general description of the cost-effectiveness of PGx testing.

OPEN

The Pharmacogenomics Journal (2015) 15, 461–466

www.nature.com/tpj

ORIGINAL ARTICLE

# Establishing the characteristics of an effective pharmacogenetic test for clozapine-induced agranulocytosis

M Verbelen[1], DA Collier[1,2], D Cohen[3], JH MacCabe[4] and CM Lewis[1,5]

Clozapine is the only evidence-based therapy for treatment-resistant schizophrenia, but it induces agranulocytosis, a rare but potentially fatal haematological adverse reaction, in less than 1% of users. To improve safety, the drug is subject to mandatory haematological monitoring throughout the course of treatment, which is burdensome for the patient and one of the main reasons clozapine is underused. Therefore, a pharmacogenetic test is clinically useful if it identifies a group of patients for whom the agranulocytosis risk is low enough to alleviate monitoring requirements. Assuming a genotypic marker stratifies patients into a high-risk and a low-risk group, we explore the relationship between test sensitivity, group size and agranulocytosis risk. High sensitivity minimizes the agranulocytosis risk in the low-risk group and is essential for clinical utility, in particular in combination with a small high-risk group.

*The Pharmacogenomics Journal* (2015) **15,** 461–466; doi:10.1038/tpj.2015.5; published online 24 February 2015

## INTRODUCTION

Although there are 22 FDA-approved antipsychotics used to treat schizophrenia, around 30% of schizophrenia patients do not respond to drugs other than clozapine.[1] Clozapine has superior efficacy for positive symptoms in these treatment-resistant patients[2] and may improve negative symptoms.[3] Furthermore, clozapine reduces suicidal behaviour especially when compared with first generation antipsychotics and overall mortality at population level.[4–8]

Despite its proven efficacy, the clinical use of clozapine is limited by the risk of agranulocytosis, a rare but potentially fatal adverse drug reaction, characterized by the acute loss of neutrophils in circulating blood. Agranulocytosis is defined as an absolute neutrophil count (ANC) of less than 500 cells $mm^{-3}$ blood. Shortly after clozapine was introduced in Europe in the 1970s, it was withdrawn from the market when 17 cases of agranulocytosis were reported in Finland, of which 8 were fatal.[9] In 1990, clozapine was reintroduced after its superiority over chlorpromazine for the treatment of refractory schizophrenia was shown.[10] However, its use was restricted in most Western countries to treatment of refractory patients, that is, patients who have not improved on at least two different antipsychotics.[2,11–13]

To prevent agranulocytosis by detecting a fall in ANC, patients treated with clozapine are subject to compulsory haematological monitoring. In Europe, the full white blood cell count and ANC are monitored weekly for the first 18 weeks of treatment and every 4 weeks thereafter for the duration of the treatment.[14] If at any time during treatment the white blood cell count falls below 3000 cells $mm^{-3}$ or the ANC below 1500 cells $mm^{-3}$, clozapine should be discontinued immediately and these patients should not be treated with clozapine again except in a controlled setting.[15–17] Although the obligatory monitoring has the benefit of regular contact with a health-care professional, it is an invasive procedure and can be a burden for the patient. Moreover, some patients decline to take clozapine because of the monitoring requirement.[18]

As agranulocytosis can develop within 2–5 days, even weekly monitoring cannot guarantee timely detection in all cases.[19] The incidence of agranulocytosis induced by clozapine varies between 0.38 and 0.8%, with approximately 80% of cases occurring within the first 18 weeks.[20–23] The incidence of agranulocytosis decreases from 0.7% in the first year, to 0.07% or lower in the second year of treatment.[24,25] Few cases occur later in the course of treatment, but the risk does not fully disappear. In 2–4% of patients, agranulocytosis is fatal, which corresponds to an overall mortality rate of about 1–3 in 10 000 patients on clozapine.[26] However, most patients recover completely from agranulocytosis with no haematological consequences.[23,24,27]

In spite of its therapeutic advantages with respect to its efficacy in treatment-resistant schizophrenia, clozapine is underused, mainly owing to the risk of severe adverse events, primarily agranulocytosis and the mandatory haematological monitoring.[28] Around 30% of schizophrenia patients meet the indications for clozapine treatment, but the market share of clozapine, which is now a generic drug, was less than 5% in 2010 in the US.[2]

A pharmacogenetic test for clozapine-induced agranulocytosis could greatly improve the burden of haematological monitoring if the monitoring requirements could be made less onerous, or be time-limited, for the majority of patients with a low genetic risk for agranulocytosis. Not only would this make clozapine treatment more acceptable for the patient, it would also save considerable health-care resources. On the other hand, the patients who are at a higher risk of developing agranulocytosis could be monitored more frequently or, if the risk is very high, not exposed to clozapine at all.

[1]SGDP Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK; [2]Discovery Neuroscience Research, Eli Lilly and Company Ltd, Lilly Research Laboratories, Erl Wood Manor, Surrey, UK; [3]Department of Severe Mental Illness, Mental Health Care Organization North-Holland North, Heerhugowaard, The Netherlands; [4]Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK and [5]Department of Medical and Molecular Genetics, King's College London, London, UK. Correspondence: Professor CM Lewis, SGDP Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London, SE5 8AF, UK.
E-mail: cathryn.lewis@kcl.ac.uk
Received 23 September 2014; revised 18 November 2014; accepted 19 December 2014; published online 24 February 2015

Pharmacogenetic research of clozapine-induced agranulocytosis has focused on candidate genes in case–control studies. Several associations with human leukocyte antigen (HLA) alleles have been reported, as well as associations with the tumour necrosis factor and N-ribosyldihydronicotin-amide quinone oxidoreductase 2 (NQO2) genes.[29,30] However, few of these findings have been replicated, and the majority of these pharmacogenetic studies suffered from typical candidate gene study issues, namely small sample sizes and inadequate correction for multiple testing. The most promising finding was that the HLA-DQB1 6672G>C polymorphism was associated with clozapine-induced agranulocytosis, with an odds ratio of 16.9.[31] A pharmacogenetic test based on this polymorphism has been marketed, but owing to low sensitivity (21.5%), it failed to be a commercial or clinical success.[29,32] In the first genome-wide association study, amino acid changes in HLA-DQB1 (126Q) and HLA-B (158T) were associated with clozapine-induced agranulocytosis with more modest odds ratios of 0.19 and 3.11, respectively.[33]

Here, we investigate the required properties of a clinically useful pharmacogenetic test that could stratify clozapine users with regards to their agranulocytosis risk as described above.

## METHODS

We assume that the genetic test divides patients into two groups with different levels of agranulocytosis risk, and that the low-risk (LR) group contains a higher proportion of patients than the high-risk (HR) group. When comparing the outcome of a pharmacogenetic agranulocytosis test with the actual agranulocytosis status, the following scenarios can occur (Table 1):

- True positive: A patient who does develop agranulocytosis is correctly identified as HR. This scenario has probability *a*.
- False positive: A patient who does not develop agranulocytosis is wrongly identified as HR. This scenario has probability *b*.
- False negative: A patient who does develop agranulocytosis is wrongly identified as LR. This scenario has probability *c*.
- True negative: A patient who does not develop agranulocytosis is correctly identified as LR. This scenario has probability *d*.

If the incidence of clozapine-induced agranulocytosis is a known parameter *k*, the agranulocytosis risk regardless of test outcome (*a*+*c*) equals *k* and the probability of not getting agranulocytosis (*b*+*d*) is 1-*k*.

Assuming that *k* is known, two of the following parameters need to be fixed to calculate the probabilities in each cell of Table 1:

- The proportion of patients in the HR group (*x*) or the proportion of patients in the LR group (1-*x*).
- The sensitivity of the test, which is the proportion of correctly classified agranulocytosis cases. In Table 1, $Sensitivity = \frac{a}{a+c} = \frac{a}{k}$
- The specificity of the test, which is the proportion of correctly classified agranulocytosis-free patients. In Table 1, $Specificity = \frac{d}{b+d} = \frac{d}{1-k}$

The proportion of patients in each risk group (*x*, 1-*x*) is relevant to this study, as we want to justify a more lenient monitoring schedule for the LR group. The larger the LR group, the more patients on clozapine will benefit from monitoring regime changes. Assuming a single locus test, the size of the two risk groups depends on the allele frequency of the test marker. Hence, we use the proportion of patients in the HR group (*x*) and test sensitivity (*s*) to study the cell probabilities in Table 1.

**Table 1.** Classification table comparing test outcome with true agranulocytosis status

| *Test result* | *Agranulocytosis* | | *Total* |
|---|---|---|---|
| | *Yes* | *No* | |
| Positive: High risk | $a = sk$ | $b = x\text{-}sk$ | $a+b = x$ |
| Negative: Low risk | $c = (1\text{-}s)\ k$ | $d = 1\text{-}x\text{-}(1\text{-}s)\ k$ | $c+d = 1\text{-}x$ |
| Total | $a+c = k$ | $b+d = 1\text{-}k$ | 1 |

Cell probabilities are expressed in terms of sensitivity (*s*) and size of the HR group (*x*).

Of primary interest is the agranulocytosis risk in the LR group, as these are the patients for whom the haematological monitoring rules could be relaxed, and this outcome corresponds to the complement of the negative predictive value (NPV), being the proportion of test negative or LR patients who do not develop agranulocytosis. Therefore, we investigate the relationship between agranulocytosis risk in the LR group, test sensitivity and the size of the HR group. The agranulocytosis risk in the LR group is given by

$$P(A|LR) = \frac{P(A \cap LR)}{P(LR)} = \frac{c}{c+d} = \frac{(1-s)k}{1-x} = 1 - NPV$$

where *A* stands for developing agranulocytosis. Thus, the agranulocytosis risk in the LR group decreases as test sensitivity increases or as the HR group becomes smaller, which corresponds to the LR group getting larger.

As a secondary outcome, we study the agranulocytosis risk in the HR group, which corresponds to the positive predictive value (PPV) or the proportion of test positive patients who are true agranulocytosis cases, and is given by

$$P(A|HR) = \frac{P(A \cap HR)}{P(HR)} = \frac{a}{a+b} = \frac{sk}{x} = PPV$$

In the HR group, the agranulocytosis risk increases when sensitivity rises or the proportion of patients in the HR group decreases.

By definition, the agranulocytosis risk in the HR group must be larger than in the LR group, so

$$P(A|HR) > P(A|LR)$$

or expressed in terms of sensitivity and size of the HR group

$$\frac{sk}{x} > \frac{(1-s)k}{1-x}$$

which reduces to $s > x$. In other words, sensitivity must be larger than the proportion of patients in the HR group.

We explore the relationship between different parameters, focusing on sensitivity (*s*), and the proportion of patients assigned to the HR group (*x*), in the context of a genetic test to predict the risk of clozapine-induced agranulocytosis. Furthermore, we develop guidelines for a test that divides the population into a LR and HR group, assuming the total risk of developing clozapine-induced agranulocytosis (*k*) is 0.8%.[11] We also assess how the pharmacogenetic test based on the HLA-DQB1 6672G>C polymorphism performs under this framework.[31]

## RESULTS

The key parameters for a clinically effective test, that is, a test that minimizes the agranulocytosis risk in the LR group, are high sensitivity and to a lesser extent a small proportion of patients assigned to the HR group. Figure 1 shows that to obtain a low agranulocytosis risk in the LR group (solid lines), with a concomitant high risk in the HR group (dotted lines), a test should be highly sensitive. Also, for a given sensitivity, a smaller HR group corresponds to lower agranulocytosis risk in the LR group and a higher agranulocytosis risk in the HR group (Figure 1 and Table 2). The lower the sensitivity of a test is, the smaller the difference between the agranulocytosis risks in both groups. When the sensitivity is equal to the size of the HR group, the risk in the two groups are the same and equal to the overall agranulocytosis risk of 0.8%. A test with sensitivity close to the proportion of patients in the HR group would thus be irrelevant.

Exploration of the relationship between agranulocytosis risk and the proportion of patients in the HR group confirms that a smaller HR group gives rise to a lower agranulocytosis risk in the LR group (Figure 2). The risk in the HR group increases steeply when the size of that group is close to zero. For a given size of the HR group, high sensitivity leads to a low agranulocytosis risk in the LR group and a high risk in the HR group (Figure 2 and Table 3). As in

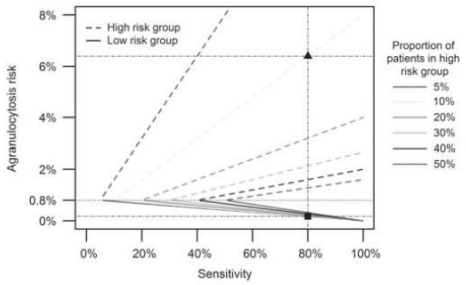



**Figure 1.** Agranulocytosis risk in LR and HR groups by sensitivity, for different HR group sizes between 5 and 50%. The ■ and ▲ indicate the agranulocytosis risk in the LR and HR groups, respectively, for a test with 80% sensitivity and 10% of patients in the HR group.

**Table 2.** Hypothetical tests with 80% sensitivity for different proportions of patients in the HR group, showing agranulocytosis risks in the LR group (1-NPV), the HR group (PPV) and specificity of the test

| *Size of HR group* | *P(A\|LR) (1-NPV)* | *P(A\|HR) (PPV)* | *Specificity* |
|---|---|---|---|
| 0.010 | 0.0016 | 0.640 | 0.996 |
| 0.050 | 0.0017 | 0.128 | 0.956 |
| 0.100 | 0.0018 | 0.064 | 0.906 |
| 0.200 | 0.0020 | 0.032 | 0.805 |
| 0.300 | 0.0023 | 0.021 | 0.704 |
| 0.400 | 0.0027 | 0.016 | 0.603 |
| 0.500 | 0.0032 | 0.013 | 0.502 |

Abbreviations: HR, high risk; LR, low risk; NPV, negative predictive value; PPV, positive predictive value.

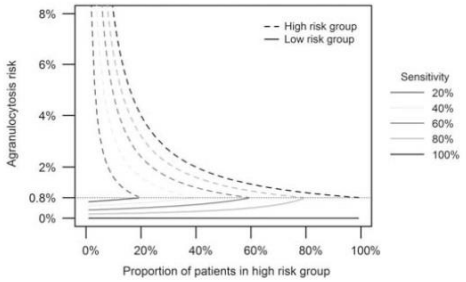



**Figure 2.** Agranulocytosis risk in LR and HR groups by proportion of patients that classified as HR, for different sensitivity values between 20 and 100%.

Figure 1, the risk curves in Figure 2 meet at 0.8% agranulocytosis risk when the proportion of patients in the HR group is equal to the sensitivity (except for 100% sensitivity where the agranulocytosis risk in the LR group is zero as all agranulocytosis cases are detected by the test).

A simultaneous assessment of sensitivity and HR group size shows that test sensitivity controls the reduction in agranulocytosis risk seen in the LR group, with high sensitivity leading to low

**Table 3.** Hypothetical tests with 10% of patients in the HR group for different sensitivity values, showing agranulocytosis risks the LR group (1-NPV), in the HR group (PPV) and specificity of the test

| *Sensitivity* | *P(A\|LR) (1-NPV)* | *P(A\|HR) (PPV)* | *Specificity* |
|---|---|---|---|
| 0.200 | 0.0071 | 0.016 | 0.901 |
| 0.400 | 0.0053 | 0.032 | 0.902 |
| 0.600 | 0.0036 | 0.048 | 0.904 |
| 0.800 | 0.0018 | 0.064 | 0.906 |
| 1 | 0 | 0.080 | 0.907 |

Abbreviations: HR, high risk; LR, low risk; NPV, negative predictive value; PPV, positive predictive value.

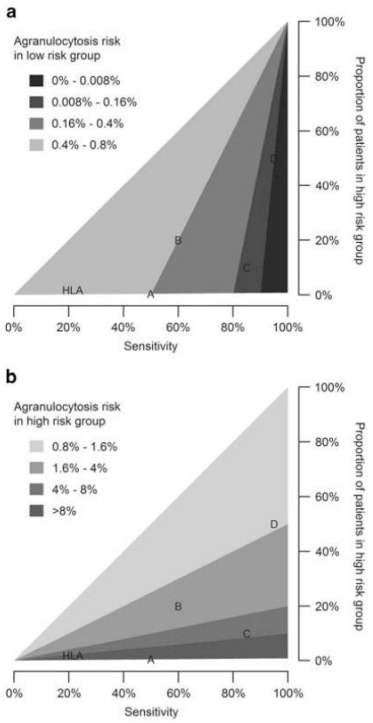



**Figure 3.** Agranulocytosis risk in (**a**) LR group and (**b**) HR group by sensitivity and proportion of patients in the HR group. In both panels, darker colours represent the desired outcomes of low risk in the LR group and high risk in the HR group. The letters indicate the position of hypothetical tests A, B, C and D; HLA indicates the position of the HLA-DQB1 6672G>C-based test.

agranulocytosis risk (Figure 3a). High sensitivity implies that most agranulocytosis cases are identified by the genetic test and classified as HR. By consequence, nearly all patients in the LR group do not develop agranulocytosis, and hence the risk in that

**Table 4.** Four hypothetical pharmacogenetic tests for clozapine-induced agranulocytosis and their clinical impact

| *Test* | *Sensitivity* | *Size of HR group* | *P(A\|LR) (1-NPV)* | *P(A\|HR) (PPV)* | *Specificity* | *Clinical impact* |
|---|---|---|---|---|---|---|
| A | 0.500 | 0.004 | 0.004 | 1 | 1 | No change in monitoring, but withhold clozapine from HR group |
| B | 0.600 | 0.200 | 0.004 | 0.024 | 0.803 | No change in monitoring |
| C | 0.850 | 0.100 | 0.0013 | 0.068 | 0.906 | Stop monitoring in LR group |
| D | 0.950 | 0.500 | 0.0008 | 0.015 | 0.504 | Stop monitoring in LR group |

Abbreviations: HR, high risk; LR, low risk; NPV, negative predictive value; PPV, positive predictive value.

group is low. For example, if we need the agranulocytosis risk in the LR group to be half the population risk, (that is, $\leqslant 0.4\%$), the test sensitivity must be at least 50.2%. To achieve a stratification where the LR group is at one-fifth of the average agranulocytosis risk, sensitivity greater than 80.1% is required. A small HR group contributes to a low agranulocytosis risk in the LR group by preventing true LR patients from being wrongly classified as HR. A large number of true LR patients maximizes the denominator of the agranulocytosis risk in the LR group, and thus minimizes the risk itself.

The agranulocytosis risk in the HR group depends largely on the size of the HR group (Figure 3b). In a smaller HR group, the ratio of true HR patients versus patients incorrectly classified as HR is larger, and so the agranulocytosis risk in the HR group is larger. High sensitivity increases the number of true HR patients and in that way leads to a high agranulocytosis risk in this group, but even in the ideal scenario of maximum sensitivity, the proportion of true HR patients is limited to 0.8%.

High sensitivity and a small HR group lead to an effective test with the small group of HR patients requiring more frequent monitoring, whereas the majority of patients are assigned to the LR group, which has substantially reduced agranulocytosis risk and could therefore be monitored less frequently.

To comment on the clinical utility of a pharmacogenetic test, an agranulocytosis risk that is acceptable without monitoring must be determined. We propose that an agranulocytosis risk of 0.13% is acceptable, because this corresponds to the risk conferred by the antipsychotic chlorpromazine which does not have mandatory monitoring in the UK.[34,35] To achieve this, the sensitivity of the test must be at least 83.9%. We examine four hypothetical pharmacogenetic tests and how the outcomes affect haematological monitoring of patients (Table 4).

- Test A is only clinically relevant for the small proportion of HR patients, but owing to its low sensitivity, the agranulocytosis risk in the LR group is too high to reduce monitoring.
- Despite a higher sensitivity than test A, test B has no definite impact on the treatment of either risk group.
- The characteristics of test C result in an agranulocytosis risk in the LR group that is low enough to stop or reduce haematological monitoring for 90% of patients. This test has the largest clinical impact.
- Although the sensitivity of test D is higher than that of test C, the larger size of the HR group in test D implies that fewer patients would benefit from this test.

The pharmacogenetic agranulocytosis test using the HLA-DQB1 6672G>C polymorphism has a sensitivity of 21.5% and specificity of 98.4%.[31] On the basis of those values and assuming an overall agranulocytosis risk of 0.8%, the test classifies 1.76% of patients as HR, with a high agranulocytosis risk of 9.66%. Conversely, the agranulocytosis risk in the LR group is 0.64%. This is a relative risk of 0.8 or a 20% reduction in risk compared with the agranulocytosis risk without genetic stratification, and exceeds our maximum acceptable agranulocytosis risk of 0.13%. Hence, the HLA-DQB1 6672G>C-based genetic test has limited use in stratifying patients in order to reduce haematological monitoring requirements for a subset of patients.

## DISCUSSION

We have established a framework for assessing the utility of a genetic test for clozapine-induced agranulocytosis and explored the characteristics of tests that would reduce agranulocytosis risk to a level that does not require regular haematological monitoring. In particular, we show that high sensitivity is essential and that a small proportion of patients classified as HR further decreases the agranulocytosis risk in the LR group.

High sensitivity is a self-evident characteristic of a clinically useful test, but the finding that a small HR group is favourable might seem counterintuitive. One could reason that to be sure the LR group contains no agranulocytosis cases, the HR group should include all patients who are at the slightest risk of developing agranulocytosis, and that the HR group should thus be large. However, as the number of true agranulocytosis cases is very low, a large HR group would mainly contain false-positive patients. Instead of a large HR group, a useful test relies on high sensitivity to correctly classify the agranulocytosis cases as HR. A small HR group implies few false positives and many true negatives, which in turn minimizes the agranulocytosis risk in the LR group.

Instead of a single genetic locus, a pharmacogenetic test could also be based on polygenic risk scores, built from combining risk conferred by many genetic loci to identify patients at high risk. In that case, the threshold defining LR and HR groups can be varied, and the effectiveness of a test is typically measured by the area under a receiver operating characteristic curve.[36] When the threshold is moved to increase test sensitivity, the size of the HR group will increase as well. It is not straightforward to predict the resulting change—increase or decrease—in agranulocytosis risks in the LR and HR groups, because these depend on the distribution of the polygenic risk scores. Once an appropriate polygenic score threshold has been fixed, a test can be translated easily to the framework developed here.

Pharmacogenetic tests for abacavir, carbamazepine and purine analogues in clinical use have sensitivity, specificity and NPV close to 100% (Table 5).[37–39] However, the PPVs of these tests vary between 7.7 and 80.6%. These low PPVs are acceptable, as there are alternative treatments available for patients who test positive. In contrast, clozapine is reserved for treatment of refractory schizophrenia and no alternative drug is available. A high PPV would ensure that few patients are unnecessarily excluded from treatment, but a high NPV and consequently low agranulocytosis risk is also important to justify a reduced monitoring schedule for patients in the LR group.

No genetic test for clozapine-induced agranulocytosis currently exists. The proposed test of HLA-DQB1 6672G>C has high specificity, but low sensitivity fails to reduce the agranulocytosis risk in the LR group sufficiently that monitoring could be reduced or ceased. Candidate gene studies have failed to identify a strong, replicated genetic variant that substantially increases risk of

**Table 5.** Characteristics of pharmacogenetic tests that predict adverse drug reactions

| *Drug* | *Gene* | *Test* | *Performance characteristics* | *Reference* |
|---|---|---|---|---|
| Abacavir | HLA-B*5701 | The HLA-B*5701 allele is associated with hypersensitivity reaction. Carriers should avoid abacavir. | Sensitivity: 100%<br>Specificity: 96.9%<br>PPV: 47.9%<br>NPV: 100% | 37 |
| Carbamazepine | HLA-B*1502 | The HLA-B*1502 allele is associated with SJS-TEN in Han Chinese. Carriers should avoid carbamazepine. | Sensitivity: 98.3%<br>Specificity: 97%<br>PPV: 7.7%<br>NPV: 100% | 38 |
| Purine analogues | TPMT | Impute 2 SNPs to identify patients with zero wildtype alleles, as they are at high risk of myelotoxicity. | Sensitivity: 100%<br>Specificity: 99.0%<br>PPV: 80.6%<br>NPV: 100% | 39 |

Abbreviations: HLA, human leukocyte antigen; NPV, negative predictive value; PPV, positive predictive value; SNP, single nucleotide polymorphism.

clozapine-induced agranulocytosis.[29,30] The first genome-wide association study of clozapine-induced agranulocytosis detected significant associations at two HLA amino acids;[33] at least one further study is in progress,[40] and combined analysis of such studies may identify associated genetic variants that can be rapidly translated to clinical practice.

## CONFLICT OF INTEREST

David A Collier is a full time employee of Eli Lilly and Company Ltd, and a visiting Professor at King's College London. David A Collier also holds stock in Eli Lilly and Company. Moira Verbelen is funded by a studentship from the Medical Research Council and Eli Lilly and Company Ltd. The remaining authors declare no conflict of interest.

## ACKNOWLEDGMENTS

This study was funded by an industrial CASE studentship to Moira Verbelen from the Medical Research Council with Eli Lilly and Company Ltd, by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 279227 (CRESTAR project, http://www.crestar-project.eu/), and under the Marie Curie Industry-Academia Partnership and Pathways, grant agreement n° 286213 (PsychDPC, http://www.psych-dpc.eu/). This study was part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

## REFERENCES

1 Meltzer HY. Treatment-resistant schizophrenia - the role of clozapine. *Curr Med Res Opin* 1997; **14**: 1–20.
2 Meltzer HY. Clozapine: balancing safety with superior antipsychotic efficacy. *Clin Schizophr Relat Psychoses* 2012; **6**: 134–144.
3 Joober R, Boksa P. Clozapine: a distinct, poorly understood and under-used molecule. *J Psychiatry Neurosci* 2010; **35**: 147.
4 Meltzer HY, Alphs L, Green AI, Altamura AC, Anand R, Bertoldi A *et al.* Clozapine treatment for suicidality in schizophrenia: international suicide prevention trial (InterSePT). *Arch Gen Psychiatry* 2003; **60**: 82–91.
5 Kiviniemi M, Suvisaari J, Koivumaa-Honkanen H, Hakkinen U, Isohanni M, Hakko H. Antipsychotics and mortality in first-onset schizophrenia: prospective Finnish register study with 5-year follow-up. *Schizophr Res* 2013; **150**: 274–280.
6 Ringback Weitoft G, Berglund M, Lindstrom EA, Nilsson M, Salmi P, Rosen M. Mortality, attempted suicide, re-hospitalisation and prescription refill for clozapine and other antipsychotics in Sweden - a register-based study. *Pharmacoepidemiol Drug Saf* 2014; **23**: 290–298.
7 Reutfors J, Bahmanyar S, Jonsson EG, Brandt L, Boden R, Ekbom A *et al.* Medication and suicide risk in schizophrenia: a nested case-control study. *Schizophr Res* 2013; **150**: 416–420.
8 Tiihonen J, Lonnqvist J, Wahlbeck K, Klaukka T, Niskanen L, Tanskanen A *et al.* 11-year follow-up of mortality in patients with schizophrenia: a population-based cohort study (FIN11 study). *Lancet* 2009; **374**: 620–627.
9 Idänpään-Heikkilä J, Alhava E, Olkinuora M, Palva IP. Agranulocytosis during treatment with chlozapine. *Eur J Clin Pharmacol* 1977; **11**: 193–198.
10 Kane J, Honigfeld G, Singer J, Meltzer H. Clozapine for the treatment-resistant schizophrenic: a double-blind comparison with chlorpromazine. *Arch Gen Psychiatry* 1988; **45**: 789.
11 Medicines and Healthcare Products Regulatory Agency *Summary of Product Characteristics: Clozaril 100mg Tablets* 2014, London. http://www.mhra.gov.uk/home/groups/spcpil/documents/spcpil/con1405056917253.pdf.
12 U.S. Food and Drug Administration *Clozaril Prescribing Information* 2013, U.S. Food and Drug Administration: Silver Spring. http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/019758s069s071lbl.pdf.
13 Warnez S, Alessi-Severini S. Clozapine: a review of clinical practice guidelines and prescribing trends. *BMC Psychiatry* 2014; **14**: 102.
14 European Agency for the Evaluation of Medicinal Products, Committed for Proprietary Medicinal Properties ( CPMP). Summary information on referral opinion following arbitration pursuant to Article 30 of Council Directive 2001/83/EC for Leponex and associated names (international non-proprietary name (INN): clozapine): Background information and Annex III. London, 2002. Retrieved from http://www.ema.europa.eu/docs/en_GB/document_library/Referrals_document/Leponex_30/WC500010966.pdf.
15 Bogers JP, Cohen D, Schulte PF, van Dijk D, Bakker B. Clozapine-induced leukopenia: arguments for rechallenge. *Ir J Med Sci* 2012; **181**: 155–156.
16 Whiskey E, Taylor D. Restarting clozapine after neutropenia: evaluating the possibilities and practicalities. *CNS Drugs* 2007; **21**: 25–35.
17 Manu P, Sarpal D, Muir O, Kane JM, Correll CU. When can patients with potentially life-threatening adverse effects be rechallenged with clozapine? A systematic review of the published literature. *Schizophr Res* 2012; **134**: 180–186.
18 Swinton M, Ahmed A. Reasons for non-prescription of clozapine in treatment-resistant schizophrenia. *Criminal Behaviour and Mental Health* 1999; **9**: 207–214.
19 Gerson SL, Meltzer H. Mechanisms of clozapine-induced agranulocytosis. *Drug Saf* 1992; **7** (Suppl 1): 17–25.
20 Honigfeld G, Arellano F, Sethi J, Bianchini A, Schein J. Reducing clozapine-related morbidity and mortality: 5 years of experience with the Clozaril National Registry. *J Clin Psychiatry* 1997; **59**: 3–7.
21 Alvir JMJ, Lieberman JA, Safferman AZ, Schwimmer JL, Schaaf JA. Clozapine-induced agranulocytosis-incidence and risk factors in the United States. *N Engl J Med* 1993; **329**: 162–167.
22 Lahdelma L, Appelberg B. Clozapine-induced agranulocytosis in Finland, 1982-2007: long-term monitoring of patients is still warranted. *J Clin Psychiatry* 2012; **73**: 837–842.
23 Munro J, O'Sullivan D, Andrews C, Arana A, Mortimer A, Kerwin R. Active monitoring of 12,760 clozapine recipients in the UK and Ireland. Beyond pharmacovigilance. *Br J Psychiatry* 1999; **175**: 576–580.
24 Atkin K, Kendall F, Gould D, Freeman H, Liberman J, O'Sullivan D. Neutropenia and agranulocytosis in patients receiving clozapine in the UK and Ireland. *Br J Psychiatry* 1996; **169**: 483–488.
25 Kumar V. Clozaril Monitoring Systems, Registry Data and Analyses (United States, United Kingdom, and Australia). *Presentation, Novartis* 2002. Accessed online on 15 September 2014 http://www.fda.gov/ohrms/dockets/ac/03/slides/3959S1_02_C-Novartis-Kumar.ppt.

26 Cohen D, Bogers JP, van Dijk D, Bakker B, Schulte PF. Beyond white blood cell monitoring: screening in the initial phase of clozapine therapy. *J Clin Psychiatry* 2012; **73**: 1307–1312.
27 Mendelowitz AJ, Gerson SL, Alvir JMJ, Lieberman JA. Clozapine-induced agranulocytosis. *CNS Drugs* 1995; **4**: 412–421.
28 Nair B, MacCabe JH. Making clozapine safer: current perspectives on improving its tolerability. *Future Neurology* 2014; **9**: 313–322.
29 Chowdhury NI, Remington G, Kennedy JL. Genetics of antipsychotic-induced side effects and agranulocytosis. *Curr Psychiatry Rep* 2011; **13**: 156–165.
30 Opgen-Rhein C, Dettling M. Clozapine-induced agranulocytosis and its genetic determinants. *Pharmacogenomics* 2008; **9**: 1101–1111.
31 Athanasiou MC, Dettling M, Cascorbi I, Mosyagin I, Salisbury BA, Pierz KA *et al.* Candidate gene analysis identifies a polymorphism in HLA-DQB1 associated with clozapine-induced agranulocytosis. *J Clin Psychiatry* 2011; **72**: 458–463.
32 Spencer BW, Prainsack B, Rujescu D, Giegling I, Collier DA, Gaughran F *et al.* Opening Pandora's box in the UK: a hypothetical pharmacogenetic test for clozapine. *Pharmacogenomics* 2013; **14**: 1907–1914.
33 Goldstein JI, Fredrik Jarskog L, Hilliard C, Alfirevic A, Duncan L, Fourches D *et al.* Clozapine-induced agranulocytosis is associated with rare HLA-DQB1 and HLA-B alleles. *Nat Commun* 2014; **5**: 4757.
34 Flanagan RJ, Dunk L. Haematological toxicity of drugs used in psychiatry. *Hum Psychopharmacol* 2008; **23** (Suppl 1): 27–41.
35 Medicines and Healthcare Products Regulatory Agency *Summary of Product Characteristics: Largactil 25mg/ml Solution for Injection* 2014, London. Retrieved from http://www.mhra.gov.uk/home/groups/spcpil/documents/spcpil/con1391143096918.pdf.
36 Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 2007; **115**: 654–657.
37 Mallal S, Phillips E, Carosi G, Molina J-M, Workman C, Tomažič J *et al.* HLA-B* 5701 screening for hypersensitivity to abacavir. *N Engl J Med* 2008; **358**: 568–579.
38 Ferrell PB, McLeod HL. Carbamazepine, HLA-B* 1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. *Pharmacogenomics* 2008; **9**: 1543–1546.
39 Almoguera B, Vazquez L, Connolly JJ, Bradfield J, Sleiman P, Keating B *et al.* Imputation of TPMT defective alleles for the identification of patients with high-risk phenotypes. *Front Genet* 2014; **5**: 96.
40 CRESTAR. CRESTAR – development of pharmacogenomic biomarkers for schizophrenia. 2011. Accessed online on 21 May 2014 http://www.crestar-project.eu/.

Editorial

Pharmacogenomics

For reprint orders, please contact: reprints@futuremedicine.com

# How close are we to a pharmacogenomic test for clozapine-induced agranulocytosis?

"The emphasis of genetic research in agranulocytosis has been on the HLA region, though other immunologic genes and clozapine metabolizing enzymes have also been investigated."

**Keywords:** agranulocytosis • clozapine • HLA • pharmacogenomic testing

Moira Verbelen
MRC SGDP Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London, SE5 8AF, UK

Cathryn M Lewis
Author for correspondence:
MRC SGDP Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London, SE5 8AF, UK
and
Department of Medical & Molecular Genetics, King's College London, 8th Floor Tower Wing, Guy's Hospital, Great Maze Pond, London, SE1 9RT, UK
Tel.: +44 20 7848 0661
Fax: +44 20 7848 0866
cathryn.lewis@kcl.ac.uk

The antipsychotic clozapine is reserved for treatment resistant schizophrenia patients despite its proven superior efficacy, because 0.8% of treated patients develop agranulocytosis, an adverse drug reaction [1–3]. Agranulocytosis is a hematological condition in which the number of neutrophils is severely reduced (less than 500 cells/mm$^3$ of blood), which renders the patient susceptible to infections and is fatal in 2–4% of agranulocytosis cases [4]. This severe adverse event is not dose dependent and few risk factors have been identified, although older age and female gender increase risk slightly [2,5]. Around 80% of agranulocytosis cases occur in the first 18 weeks of clozapine treatment and the risk reduces by a factor of ten after the 1st year [1,2]. To improve safety, patients treated with clozapine are required to undergo regular white blood cell counts throughout the course of treatment. This is an invasive procedure, which may be stressful and inconvenient for the patients, and is an important factor in explaining why clozapine is underused despite its efficacy. The market share of clozapine in the USA was only 5% in 2008, whereas around 40% of patients were eligible [3].

A pharmacogenomic test for clozapine-induced agranulocytosis could be used for two different purposes. A test that identifies patients at very low risk would allow some patients to be exempted from hematological monitoring or to follow a more relaxed monitoring schedule. This should remove a barrier for clozapine usage and increase clinical uptake. Alternatively, a test that picks up patients who are at increased risk could be used to improve safety by ensuring these high-risk patients are monitored more closely or not prescribed clozapine. Keeping in mind that there is no alternative treatment for clozapine, the number of false positive test results should be absolutely minimal. Both applications of genetic tests are important, but a test that identifies the low-risk patients has a potential impact on a much larger group of patients and is therefore our primary focus.

## Pharmacogenomics of clozapine-induced agranulocytosis

The emphasis of genetic research in agranulocytosis has been on the HLA region, though other immunologic genes and clozapine metabolizing enzymes have also been investigated. From the 1990s, several candidate gene studies reported associations with HLA-alleles and haplotypes, but few results have been replicated [6,7]. This limited success can partly be explained by the small sample sizes of these studies: around 30 agranulocytosis cases gives only limited power to detect an association. Even more important to note is that many of these statistical findings were not adjusted for multiple testing, thereby increasing the probability of false positive results. The most promising candidate gene study found an association with the SNP *HLA-DQB1* 6672G>C and replicated this finding in a separate sample, using 82 cases

Future Medicine part of fsg

in total [8]. The odds ratio (OR) of this association was 17, with 21.5% sensitivity and 98.4% specificity.

> "Despite this growing evidence that the HLA region is associated with agranulocytosis in clozapine patients, this knowledge has not been translated to clinical practice yet."

More recently, researchers have broadened the search for genetic associations with clozapine-induced agranulocytosis beyond candidate genes. Exome sequencing in a small Finnish sample investigated SNPs, insertions and deletions but found no significant associations [9]. Goldstein *et al.* performed the most extensive genetic study to date encompassing a genome-wide association study, rare variant exome sequencing, copy number variants analysis and HLA allele imputation in 163 agranulocytosis cases [10]. Two amino acid changes in *HLA-DQB1* 126Q (OR = 0.19) and *HLA-B* 158T (OR = 3.11) reached genome-wide significance levels. Combining these two HLA alleles in a pharmacogenomic test that classifies a patient carrying either variant as being at high risk of agranulocytosis would achieve 36% sensitivity and 89% specificity. However, the authors comment that "these ORs do not immediately suggest clinical application in screening" [10].

### Application of pharmacogenomic findings to clinical practice

Despite this growing evidence that the HLA region is associated with agranulocytosis in clozapine patients, this knowledge has not been translated to clinical practice yet. In 2007, a test based on the *HLA-DQB1* 6672G>C variant was commercially marketed but its low sensitivity meant that only one in five agranulocytosis patients were detected and the risk among the patients classified as low risk by the test was only reduced by 20%. Consequently, the test was not widely used and is no longer available [11,12].

> "A further puzzle: assuming our current knowledge of the genetic underpinnings of clozapine-induced agranulocytosis is incomplete, where do the remaining risk variants lie?"

We recently developed a framework to assess the clinical value of pharmacogenomic tests for clozapine-induced agranulocytosis and concluded that sensitivity is the most important parameter for evaluating clinical utility [13]. High sensitivity ensures that patients who are at high risk are correctly identified by the test, and that therefore few patients considered low risk will develop agranulocytosis. The necessity of high sensitivity explains why the *HLA-DQB1* 6672G>C test failed commercially and why the amino acid changes in *HLA-DQB1* and *HLA-B* are not suitable for predictive testing on their own.

Looking at the whole field of psychiatric drug treatment, we must concede that clinical application of genomic findings is still modest. The US FDA lists 24 psychiatric drugs (including clozapine) with pharmacogenomic information on their drug label; in all cases this relates to cytochrome P450 metabolism and possible consequences on drug dosing [14]. However, pharmacogenomic testing is required by the FDA for only one of these 24 drugs, the antipsychotic pimozide [15]. A test for clozapine-induced agranulocytosis could be the first test for an adverse drug reaction caused by a psychiatric drug.

### Challenges of identifying pharmacogenomic test for clozapine-induced agranulocytosis

The monitoring procedures introduced for clozapine have been successful in reducing the number of cases of clozapine-induced agranulocytosis, but this clinical success makes it difficult to attain the large sample size of agranulocytosis cases required for well-powered genetic studies. One alternative would be to use patients with neutropenia given a 'red flag' in monitoring (neutrophil count less than 1500 cells/mm$^3$) or even include those with an 'amber flag' (neutrophil count 1500–2000 cells/mm$^3$). However, relaxing the case threshold for ascertainment is usually counterproductive in genetic studies as patients who would have developed agranulocytosis may be outnumbered by those whose neutrophil levels dropped for other reasons. A further puzzle: assuming our current knowledge of the genetic underpinnings of clozapine-induced agranulocytosis is incomplete, where do the remaining risk variants lie? Methods to impute HLA alleles from SNP genotypes have substantially expanded our ability to assess this type of variation in large genetic studies [16]. In addition to *HLA-DQB1* 126Q and *HLA-B* 158T, lower frequency or lower risk HLA variants may exist, which this study was underpowered to detect, and this is suggested by residual evidence for association in the HLA region [10]. Variants outside the HLA region may also increase agranulocytosis risk, and new genome-wide association studies such as the EU-funded CRESTAR study are in progress [17].

## Conclusion

Candidate gene studies have targeted the HLA region and the largest genome-wide study so far has indeed confirmed that amino acid changes in HLA alleles are associated with clozapine-induced agranulocytosis. Hypothesis free genome-wide association studies have proven to be more successful than candidate gene studies in disease genetics and larger studies investigating

the whole genome should also enable us to uncover more pharmacogenomic associations. In addition, meta-analyses and multidisciplinary collaboration between researchers worldwide will be essential in progressing this field. If genetic variation controls risk of clozapine-induced agranulocytosis, we now have the tools and technology to identify the variants, and to translate them into a highly sensitive pharmacogenomic test that will improve safety, clinical care and patient experience.

### Financial & competing interests disclosure

This study was funded by an industrial CASE studentship to M Verbelen from the Medical Research Council with Eli Lilly and Company Ltd, by the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 279227 (CRESTAR project, www.crestar-project.eu), and under the Marie Curie Industry–Academia Partnership and Pathways, grant agreement n° 286213 (PsychDPC, www.psych-dpc.eu). This study was part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

## References

1 Munro J, O'Sullivan D, Andrews C, Arana A, Mortimer A, Kerwin R. Active monitoring of 12,760 clozapine recipients in the UK and Ireland. Beyond pharmacovigilance. *Br. J. Psychiatry* 175(6), 576–580 (1999).

2 Atkin K, Kendall F, Gould D, Freeman H, Liberman J, O'Sullivan D. Neutropenia and agranulocytosis in patients receiving clozapine in the UK and Ireland. *Br. J. Psychiatry* 169(4), 483–488 (1996).

3 Meltzer HY. Clozapine: balancing safety with superior antipsychotic efficacy. *Clinical Schizophrenia & Related Psychoses* 6(3), 134–144 (2012).

4 Cohen D, Bogers JP, Van Dijk D, Bakker B, Schulte PF. Beyond white blood cell monitoring: screening in the initial phase of clozapine therapy. *J. Clin. Psychiatry* 73(10), 1307–1312 (2012).

5 Alvir JMJ, Lieberman JA, Safferman AZ, Schwimmer JL, Schaaf JA. Clozapine-induced agranulocytosis--incidence and risk factors in the United States. *N. Engl. J. Med.* 329(3), 162–167 (1993).

6 Opgen-Rhein C, Dettling M. Clozapine-induced agranulocytosis and its genetic determinants. *Pharmacogenomics* 9(8), 1101–1111 (2008).

7 Chowdhury NI, Remington G, Kennedy JL. Genetics of antipsychotic-induced side effects and agranulocytosis. *Current Psychiatry Reports* 13(2), 156–165 (2011).

8 Athanasiou MC, Dettling M, Cascorbi I *et al.* Candidate gene analysis identifies a polymorphism in HLA-DQB1 associated with clozapine-induced agranulocytosis. *J. Clin. Psychiatry* 72(4), 458–463 (2011).

9 Tiwari AK, Need AC, Lohoff FW *et al.* Exome sequence analysis of Finnish patients with clozapine-induced agranulocytosis. *Mol. Psychiatry* 19(4), 403–405 (2014).

10 Goldstein JI, Fredrik Jarskog L, Hilliard C *et al.* Clozapine-induced agranulocytosis is associated with rare HLA-DQB1 and HLA-B alleles. *Nature Communications* 5, 4757 (2014).

11 Spencer BW, Prainsack B, Rujescu D *et al.* Opening pandora's box in the UK: a hypothetical pharmacogenetic test for clozapine. *Pharmacogenomics* 14(15), 1907–1914 (2013).

12 Pouget JG, Shams TA, Tiwari AK, Müller DJ. Pharmacogenetics and outcome with antipsychotic drugs. *Dialogues Clin. Neurosci.* 16(4), 555–566 (2014).

13 Verbelen M, Collier DA, Cohen D, Maccabe JH, Lewis CM. Establishing the characteristics of an effective pharmacogenetic test for clozapine-induced agranulocytosis. *Pharmacogenomics J.* (2015).

14 US Food and Drug Administration. Table of pharmacogenomic biomarkers in drug labeling (2015). www.fda.gov

15 Pharmgkb. Drug labels (2015). www.pharmgkb.org/view/drug-labels.do

16 Jia X, Han B, Onengut-Gumuscu S *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* 8(6), e64683 (2013).

17 Crestar. Crestar – development of pharmacogenomic biomarkers for schizophrenia (2015). www.crestar-project.eu

OPEN

The Pharmacogenomics Journal (2017) 00, 1–8

www.nature.com/tpj

REVIEW

# Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet?

M Verbelen[1], ME Weale[2] and CM Lewis[1,2]

Pharmacogenetics (PGx) has the potential to personalize pharmaceutical treatments. Many relevant gene–drug associations have been discovered, but PGx-guided treatment needs to be cost-effective as well as clinically beneficial to be incorporated into standard health-care. We reviewed economic evaluations for PGx associations listed in the US Food and Drug Administration (FDA) Table of Pharmacogenomic Biomarkers in Drug Labeling. We determined the proportion of evaluations that found PGx-guided treatment to be cost-effective or dominant over the alternative strategies, and estimated the impact on this proportion of removing the cost of genetic testing. Of the 137 PGx associations in the FDA table, 44 economic evaluations, relating to 10 drugs, were identified. Of these evaluations, 57% drew conclusions in favour of PGx testing, of which 30% were cost-effective and 27% were dominant (cost-saving). If genetic information was freely available, 75% of economic evaluations would support PGx-guided treatment, of which 25% would be cost-effective and 50% would be dominant. Thus, PGx-guided treatment can be a cost-effective and even a cost-saving strategy. Having genetic information readily available in the clinical health record is a realistic future prospect, and would make more genetic tests economically worthwhile.

*The Pharmacogenomics Journal* advance online publication, 13 June 2017; doi:10.1038/tpj.2017.21

## INTRODUCTION

Pharmacogenetics (PGx) studies the relationship between genetic variation and inter-individual variability in drug response in terms of efficacy and safety. Hence, PGx knowledge can be used to tailor pharmaceutical treatment to the genetic make-up of the patient. Several robust, well-replicated PGx associations exist, for example, the association of *HLA-B*5701* with abacavir hypersensitivity, *HLA-B*1502* with carbamazepine-induced Stevens–Johnson syndrome/toxic epidermal necrolysis, and *VKORC1* and *CYP2C9* with warfarin dosing.[1–3] Accordingly, the US Food and Drug Administration (FDA) includes information about PGx associations in many drug labels in a wide range of therapeutic areas.[4] These PGx drug labels cover tests that are commonly used, but also include weaker genetic associations that are reported without requiring adjustments to pharmaceutical treatment. Most drugs with mandatory genetic testing are used in oncology, but PGx tests in other therapeutic areas are already being offered by laboratories and some have become part of standard clinical practice.[5,6]

Personalizing drug treatments through PGx testing could improve their efficacy and safety, as well as reduce costs.[7] However, as health-care resources are finite, it is important that the cost-effectiveness of novel PGx-guided treatment strategies is assessed in addition to their clinical utility before they are widely applied. Economic evaluations, which compare costs and outcomes of at least two competing interventions, are a useful tool to inform decision making and prioritize health-care spending. In the context of PGx testing, a pharmaco-economic study might contrast PGx-guided treatment with standard treatment (ST) with the same drug, or with an alternative drug that does not require genetic testing, or with both alternatives. When the PGx strategy is found to be more effective at an acceptable additional cost (cost-effective) or more effective at a lower cost (cost-saving or dominant), this provides a strong argument for the implementation of PGx testing.

Previously published literature reviews of PGx-guided treatment and personalized medicine reported that the majority of PGx strategies were cost-effective or even dominant, though they noted that there was large heterogeneity in methodology between studies.[8–12] Concerns over the quality of the early economic evaluations of PGx-guided treatment have been raised, but the quality is generally considered to have improved over time.[13–16]

Our review of pharmaco-economic studies of PGx-guided treatment provides an update on the literature in this rapidly evolving field (the most recent previous review covered studies up to early 2013 (ref. 10)). Furthermore, we include a more extensive range of economic evaluations, whereas recent literature reviews were limited to cost utility analyses (CUAs) only.[10,11] We also assessed the impact of freely available genetic information on the cost-effectiveness of PGx-guided treatment. We adopted a narrow definition of PGx, limiting our scope to consideration of variation in germline DNA. In contrast to tests on tumour, viral or bacterial DNA, germline DNA has the advantage that genetic variants need to be typed only once, and results remain relevant throughout a patient's life.

## MATERIALS AND METHODS

### Data sources and search strategy

The FDA Table of Pharmacogenomic Biomarkers in Drug Labeling lists FDA-approved drugs that include PGx information on their

[1]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK and [2]Division of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, UK. Correspondence: Professor CM Lewis, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK.
E-mail: cathryn.lewis@kcl.ac.uk
Received 21 September 2016; revised 15 February 2017; accepted 14 April 2017

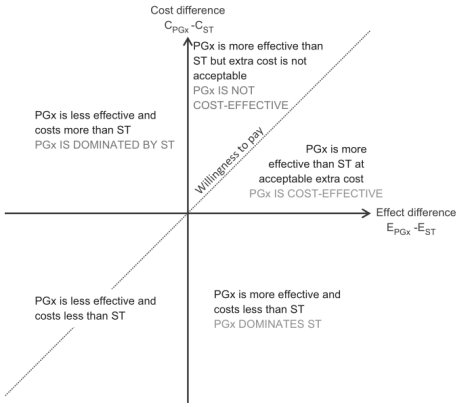



**Figure 1.** Cost-effectiveness plane of pharmaco-economic studies. PGx, pharmacogenetics-guided treatment; ST, standard treatment.

drug label along with the biomarker gene (accessed on 18 September 2015).[4] We used this table to identify drugs for which there is a genetic variant associated with the drug efficacy, safety or dosing. We excluded non-germline genetic biomarkers, for example, mutations in viral or tumour DNA.

We then searched for the selected drugs and biomarkers in the National Health Service Economic Evaluations Database (NHS EED), a UK Department of Health and National Institute for Health Research-funded registry of economic evaluations of health and social care interventions.[17,18] This resource includes CUAs, cost-effectiveness analyses (CEAs), cost–benefit analyses (CBAs—see below for definitions of these terms) and commentaries by the Centre for Reviews and Dissemination of the University of York. Funding of the NHS EED ceased in March 2015 and the latest database update was December 2014.

For each drug included in our study, the NHS EED was searched for economic evaluations that contain (1) the drug name and (2) the specific gene from the FDA label or the search terms genetic, genotype, genotypic, pharmacogenetic or pharmacogenomic in any field. We only included studies that compared a PGx-guided treatment strategy with at least one alternative strategy.

We also searched PubMed to identify more recent papers (until September 2015) and any other studies missed by the NHS EED search. We searched for articles that included (1) the name of the drug and (2) the specific gene mentioned in the FDA label or the search terms genetic, genotype, genotypic, pharmacogenetic or pharmacogenomic in the title or abstract, and (3) Cost-Benefit Analysis as a Medical Subject Headings term. In addition, the reference lists of retrieved publications were used to identify additional studies missed in our database searches.

### Overview of economic evaluation methodology

Measuring and comparing costs and health outcomes is essential in a pharmaco-economic study. Whereas costs are naturally expressed in monetary units, the effect of a healthcare intervention can be expressed in different ways. In CUAs, health outcomes are assessed as quality-adjusted life years (QALYs), which measure the expected number of post-treatment years of life accounting for the quality of life. QALYs allow comparisons of treatment strategies across therapeutic areas and populations, but are an abstract concept ('quality' is hard to define) and their validity has been questioned.[19] CEAs evaluate the effect of an intervention in terms of a disease or treatment specific measure, for example the number of adverse events avoided, the change in score on a depression rating scale or time taken to remission. CBAs quantify treatment outcome in purely monetary terms.

Furthermore, the perspective of a pharmaco-economic study determines which costs and benefits are taken into account. These can be limited to costs to the public health-care system or private insurers, for example, staff salaries, drugs and equipment costs, or may include broader costs such as productivity losses and informal care. Commonly used perspectives are the third-party payer and societal perspective, but some studies take a hospital or patient perspective.

The incremental cost-effectiveness ratio (ICER) summarizes the difference in costs and health outcomes between a PGx-guided strategy and ST:

$$\mathrm{ICER} = \frac{(\mathrm{Cost}_{\mathrm{PGx}} - \mathrm{Cost}_{\mathrm{ST}})}{(\mathrm{Effect}_{\mathrm{PGx}} - \mathrm{Effect}_{\mathrm{ST}})}$$

If the PGx treatment reduces costs and achieves a better outcome than the ST, then the PGx strategy dominates the ST. Contrarily, if the PGx option costs more but is less effective than the ST, then the PGx treatment is dominated by the ST (Figure 1). When one treatment comes at a higher cost but is also more effective than the other, the ICER is compared to a willingness-to-pay threshold to determine cost-effectiveness. Generally, ICERs up to £20 000–£30 000 per QALY (or $30 000–$50 000 per QALY) are considered cost-effective.[20] As costs, health outcomes and willingness-to-pay thresholds differ between countries, or may differ according to the assumptions and perspectives adopted, economic studies evaluating the same PGx test may come to different conclusions.

### Analyses

We extracted key parameters from the reviewed economic evaluations, including the unit of outcome, country, perspective, ICER if applicable and the conclusion regarding the cost-effectiveness of the PGx testing strategy (the interpretation of the result as described in the publication). A parameter of particular interest is the cost of the genetic test, as this can significantly affect the cost-effectiveness of the PGx testing strategy and may change over time. To allow comparison between studies, the price of the genetic test was corrected for inflation and converted to US dollars estimated at 2014 levels (2014 US$).

A stepwise linear regression model was fitted to test whether publication year, geographic region (Asia, Oceania, United States and Canada or EU) or perspective (health care, society or other) had an influence on the price of genetic testing. A stepwise logistic regression model was also used to investigate whether publication year, geographic region, perspective, cost of genetic test, genetic variant (*HLA*, *TPMT* or other) or outcome (QALY or other) was associated with the PGx testing strategy being cost-effective. Statistical analysis was performed in R (version 3.1.2, R Foundation for Statistical Computing, Vienna, Austria).

We estimated the impact of freely available genetic information on the conclusions regarding the cost-effectiveness of PGx-informed strategies. The ICER under assumption of free genetic testing was calculated by adjusting the cost of the PGx-guided treatment for the cost of the test as reported in the reviewed studies

$$\mathrm{ICER}_{\mathrm{free\ PGx}} = \frac{\left[\mathrm{Cost}_{\mathrm{PGx}} - \mathrm{Cost}_{\mathrm{genetic\ test}} - \mathrm{Cost}_{\mathrm{ST}}\right]}{(\mathrm{Effect}_{\mathrm{PGx}} - \mathrm{Effect}_{\mathrm{ST}})}$$

When insufficient details were provided to estimate the $\mathrm{ICER}_{\mathrm{free\ PGx}}$, it was assumed that free genetic testing could not worsen the conclusion regarding PGx-guided treatment. For example, when a study found the PGx strategy to be cost-effective, we assumed that PGx-guided treatment with free genetic testing would also be at least as cost-effective.

## RESULTS

### Description of studies

The FDA Table of Pharmacogenomic Biomarkers in Drug Labeling listed 137 distinct drugs, of which 68 met our inclusion criteria (Figure 2). These drugs were from diverse clinical specialties, including cancer (11 drugs), infectious diseases (10 drugs), psychiatry (9 drugs) and neurology (8 drugs) (Table 1). Our literature search yielded economic evaluations for only 10 of these 68 drugs (14.7%; Table 2). All publications related to a single drug, except for one study investigating a PGx testing strategy for carbamazepine and phenytoin treatment, which assumed both drugs to be interchangeable in terms of costs, efficacy and safety.[21] To avoid duplication of studies in our review, this publication was counted as a carbamazepine study (there were no other publications on phenytoin).

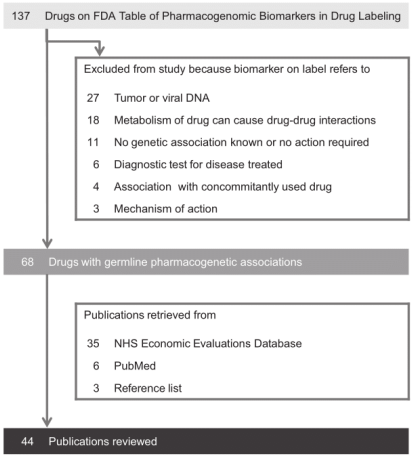



**Figure 2.** Number of drugs and publications included in literature review.

We retrieved 44 economic evaluations that investigated the cost-effectiveness of a PGx-informed strategy (Table 2). Full details of the reviewed studies and extracted information are given in Supplementary Table 1. The earliest study included was published in 2000 and over 70% of studies were published in 2009 or later. Most publications were CUAs (30 studies, 68%) or CEAs (12 studies, 27%), with only two CBAs (5%). A health-care system perspective was adopted in 18 studies (41%), a societal perspective in 10 papers (23%), a third-party payer perspective in 5 studies (11%) and 11 papers (25%) did not state a clear perspective. In all, 20 studies (45%) were conducted in North America, 11 (25%) in Europe, 6 (14%) in Asia and 3 (7%) in Oceania; 4 studies (9%) did not specify a country. Warfarin had the most economic evaluations (12 studies), followed by azathioprine (9 studies); clozapine and mercaptopurine had only 1 economic evaluation each (Table 2).

### Cost-effectiveness of PGx-informed treatment

We assessed the overall conclusions regarding cost-effectiveness of each PGx study. Over half of the 44 economic evaluations took a favourable view of the PGx-guided strategy: in 12 studies (27%) it was dominant (cost-saving) and in 13 studies (30%) it was cost-effective. Eleven publications (25%) found PGx testing not cost-effective and 8 studies (18%) did not reach a definitive conclusion (Figure 3a). The majority of economic evaluations concluded in favour of PGx testing for azathioprine (7 out of 9 studies), clopidogrel (4 out of 6 studies), abacavir (4 out of 5 studies), carbamazepine (3 out of 4 studies), irinotecan (3 out of 3 studies) and clozapine (1 study) (Figure 3b). Although warfarin had the highest number of economic studies, they reached diverging conclusions: 3 studies found PGx-guided dosing cost-effective, 4 studies were inconclusive and 5 studies concluded it was not cost-effective. No studies found unequivocally that PGx-guided

**Table 1.** Drugs from the FDA Table of Pharmacogenomic Biomarkers in Drug Labeling included in literature review

| *Therapeutic area* | *Count* | *Drugs* |
|---|---|---|
| Oncology | 11 | Capecitabine, cisplatin, dabrafenib, fluorouracil, **irinotecan**, lapatinib, **mercaptopurine**, nilotinib, pazopanib, rasburicase, thioguanine |
| Infectious diseases | 10 | **Abacavir**, chloroquine, dapsone, mafenide, nalidixic acid, nitrofurantoin, primaquine, quinine sulphate, rifampin+isoniazid+pyrazinamide[a], sulfamethoxazole+trimethoprim[a] |
| Psychiatry | 9 | Aripiprazole, atomoxetine, **citalopram**, **clozapine**, fluvoxamine, iloperidone, perphenazine, pimozide, thioridazine |
| Neurology | 8 | **Carbamazepine**, clobazam, dextromethorphan+quinidine[a], divalproex, **phenytoin**, tetrabenazine, valproic acid, vortioxetine |
| Cardiology | 5 | Carvedilol, **clopidogrel**, isosorbide+hydralazine[a], metoprolol, propafenone |
| Gastroenterology | 5 | Dexlansoprazole, esomeprazole, metoclopramide, PEG-3350+sodium sulphate+sodium chloride+potassium chloride+sodium ascorbate+ascorbic acid[a], rabeprazole |
| Rheumatology | 5 | **Azathioprine**, carisoprodol, celecoxib, flurbiprofen, pegloticase |
| Endocrinology | 4 | Chlorpropamide, glimepiride, glipizide, glyburide |
| Haematology | 3 | Eltrombopag, methylene blue, **warfarin** |
| Analgesic | 1 | Tramadol |
| Anaesthesiology | 1 | Codeine |
| Dental | 1 | Cevimeline |
| Genitourinary | 1 | Tolterodine |
| Inborn errors of metabolism | 1 | Eliglustat |
| Pulmonary | 1 | Ivacaftor |
| Toxicology | 1 | Sodium nitrite |
| Transplantation | 1 | Mycophenolic acid |

Abbreviation: FDA, Food and Drug Administration. Drugs in bold had economic evaluations available. [a]Multiple drugs on a single FDA label.

**Table 2.** Drugs for which a PGx-guided strategy was studied in economic evaluation(s)

| *Drug* | *Therapeutic area* | *Gene* | *Notes (based on PharmGKB.org[37])* | *Number of reviewed publications* |
|---|---|---|---|---|
| Abacavir | HIV | *HLA-B* | Abacavir is contraindicated for *HLA-B*5701* carriers as they are at high risk of hypersensitivity reaction. | 5 (refs 38–42) |
| Azathioprine | Rheumatology | *TPMT* | Carriers of one nonfunctional *TPMT* allele may require reduced azathioprine dose. Carriers of two nonfunctional *TPMT* alleles are at high risk of myelotoxicity and alternative treatment should be considered. | 9 (refs 43–51) |
| Carbamazepine | Neurology | *HLA-B, HLA-A* | Carbamazepine is contraindicated for *HLA-B*1502* carriers as they are at high risk Stevens–Johnson syndrome/toxic epidermal necrolysis. *HLA-A*3101* has also been associated with hypersensitivity reactions. | 4 (refs 21,52–54) |
| Citalopram | Psychiatry | *CYP2C19, 5-HTTLPR* [a]*, HTR2A* [a] | *CYP2C19* poor metabolizers require reduced citalopram starting dose. Polymorphisms in 5-HTTLPR and HTR2A are associated with citalopram response.[55,56] | 3 (refs 57–59) |
| Clopidogrel | Cardiology | *CYP2C19* | *CYP2C19* poor metabolizers have reduced response to clopidogrel and alternative treatment should be considered. | 6 (refs 25,26,60–63) |
| Clozapine | Psychiatry | *CYP2D6, H2* [a]*, 5-HTT* [a]*, 5-HT$_{2A}$,* $5\text{-HT}_{2C}$[a] | *CYP2D6* poor metabolizers may require reduced clozapine dose. Six polymorphisms in H2, 5-HTT, 5-HT2A and 5-HT2C are associated with clozapine response.[64] | 1 (ref. 65) |
| Irinotecan | Oncology | *UGT1A1* | Patients homozygous for the *UGT1A1*28* allele are at higher risk of neutropenia and should receive a reduced starting dose of irinotecan. | 3 (refs 66–68) |
| Mercaptopurine | Oncology | *TPMT* | Carriers of one nonfunctional *TPMT* allele may require reduced mercaptopurine dose. Carriers of two nonfunctional *TPMT* alleles are at high risk of myelotoxicity and alternative treatment should be considered. | 1 (ref. 69) |
| Warfarin | Cardiology | *CYP2C9, VKORC1* | Genetic variation in *VKORC1* and *CYP2C9* explain 40% variance in warfarin dose. Genetic and clinical information can be used to determine starting dose. | 12 (refs 22–24,30,70–77) |

Abbreviation: PGx, pharmacogenetics. [a]Gene not mentioned on FDA drug label but appears in economic evaluations.

citalopram (3 studies) or mercaptopurine (1 study) treatment was cost-effective.

We assessed the effect of study characteristics on the probability of concluding in favour of the PGx strategy. A logistic regression model detected that CUAs (studies using QALYs as outcome measure) were less likely than CEAs and CBAs to find the genetic testing strategy cost-effective (odds ratio = 0.13, *P*-value $< 0.05$). However, there is no clear explanation for this and it may be a spurious result due to the relatively small sample size of 44 economic evaluations.

### Effect of cost of genetic test on cost-effectiveness of PGx-informed treatment

The cost of genetic testing is an important parameter of economic evaluations of PGx interventions. After correcting for inflation and converting to 2014 US$, the cost of genetic testing quoted by the reviewed studies ranged between US$33 and US$710 with a median value of US$175. The price of genetic tests decreased slightly over time (not statistically significant) and this trend was more pronounced since 2009, the period when most economic evaluations were published (*P*-value $< 0.05$; Figure 4). Prices were on average higher in the United States and Canada than other regions of the world (mean United States and Canada: US$363.65; mean other regions: US$131.80; *P*-value $< 0.05$). We noted a wide variability in prices of tests for the same drug. For example, the lowest price quoted for warfarin PGx testing was US$36 in a 2014 UK-based study,[22] while US$600 and US$657 were used in a 2013 Canadian and 2009 US study, respectively.[23,24] The prices for clopidogrel PGx testing also varied considerably: from US$45 (2013 Australian study) to US$ 543 (2013 US study).[25,26]

Given the decreasing costs of genetic testing and its increasing availability, we looked ahead to a possible future where genotype information might be readily available, at negligible cost, for all patients as part of their electronic health record. Thirty-three economic evaluations (75%) would support PGx-guided treatment under this scenario, with 11 studies (25%) finding it cost-effective and 22 studies (50%) considering it dominant and cost-saving (Figure 3c). Five studies (11%) would still conclude that PGx testing was not cost-effective, while 3 studies (7%) would be inconclusive. A separate set of 3 studies had to be excluded, because the impact of free genetic testing could not be estimated. We note that the effect of freely available genetic information can be striking for some drugs. None of the published studies for citalopram and mercaptopurine found PGx-informed treatment to be cost-effective, but all studies switched in favour of PGx testing under the negligible test cost scenario (Figure 3d). For the 12 economic evaluations of warfarin, the number of cost-effective studies would increase from 3 to 7 with freely available genetic testing.

## DISCUSSION

We have assessed published economic evaluations comparing the cost-effectiveness of PGx-guided treatment to ST for drugs listed in the FDA Table of Pharmacogenomic Biomarkers in Drug Labeling. The economic evaluations were drawn from the NHS EED database, which includes economic evaluations up to 31 December 2014. An alternative source of economic studies would be the Cost-Effectiveness Registry (CEA Registry) maintained by the Tufts Medical Centre. We opted to use the more comprehensive NHS EED as the CEA Registry is limited to CUAs (measuring health outcomes in QALYs), which would have reduced the number of evaluations available for assessment. Moreover, the CEA Registry was not updated beyond 2014 and it only provides advanced database searches for subscribers and contributors.[18] A third resource, the Health Economic Evaluations Database curated by John Wiley & Sons, was discontinued in 2014.[27] As economic

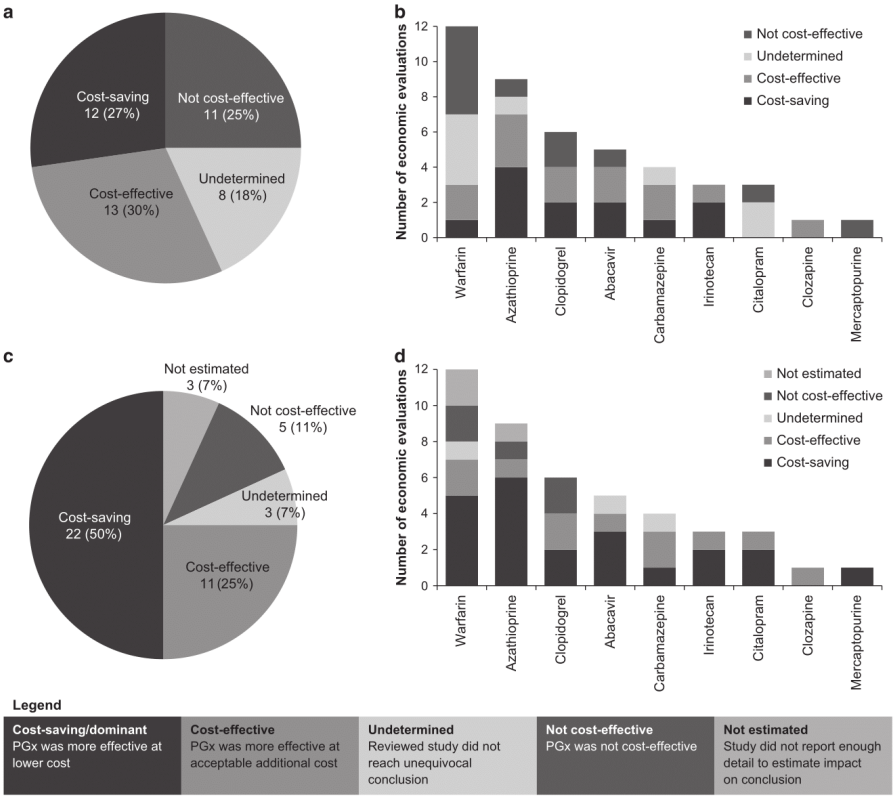



**Figure 3.** Conclusions of reviewed economic evaluations regarding cost-effectiveness of PGx testing strategy (**a**) overall and (**b**) by drug, and estimated conclusions in scenario of no extra cost for genetic information (**c**) overall and (**d**) by drug. PGx, pharmacogenetics-guided treatment.

evaluations provide evidence for the introduction of PGx testing into clinical practice, we argue that an up-to-date, accessible database would be an important and valuable resource for both health-economic and PGx research.

Few of the FDA-listed drugs have been the subject of published economic evaluations assessing the economic aspects of PGx testing. This was previously also noted by Phillips *et al.*,[11] who found that only 13% of drugs on the FDA table and only 27% of available genetic tests were accompanied by economic studies. However, it is increasingly the case that clinical utility alone is not sufficient to recommend application of a PGx test in clinical practice, and a favourable economic assessment is therefore of increasing importance. We call for more pharmaco-economic studies in this field, which should be regularly updated to respond to the changing landscape of health-care and, in particular, genetic testing costs.

There are various limitations of our study that need to be taken into account. One is that the economic evaluations reviewed may not be representative of all PGx tests. For example, the economic aspects of PGx-guided treatment are of less relevance in cases where testing is clearly necessary, for example, because it prevents life-threatening adverse events, and economic studies in such cases therefore tend to be lacking. Another possibility is that economic studies focus on PGx tests that are already applied in clinical practice and for which there is an apparent interest. Studies that find genetic testing to be not cost-effective may also be less likely to be published.

Notwithstanding the above issues, economic evaluations also have certain intrinsic limitations. One is that certain inputs into the model are difficult to quantify accurately. For example, parameters such as the response rate, the probability of adverse drug reactions and the cost of managing adverse drug reactions must sometimes be estimated from sparse information. Randomized clinical trials are the preferred source for these input data, but these are not always available. Ideally, the uncertainty in the input estimates should be accounted for in the economic modelling, and sensitivity analyses should be performed to verify how robust the result is to deviations in the inputs, but the level of uncertainty to apply can itself be a matter of subjective opinion, and vary from study to study.

Another intrinsic issue is that context and perspective may influence the conclusion of a study. For example, comparing

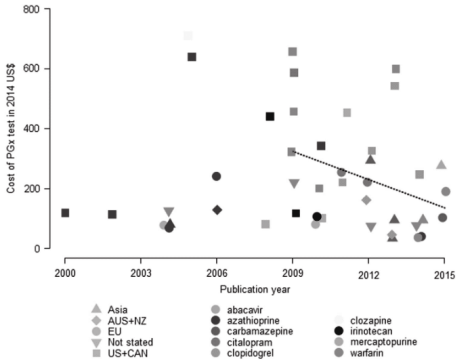



**Figure 4.** Cost of pharmacogenomics (PGx) test as reported in the reviewed economic evaluations over time, with fitted regression since 2009 (dotted line).

treatments from the perspective of an insurance company over 5 years will count costs and outcomes differently from looking at the same treatment options from a broader societal perspective. Likewise, economic evaluations are typically country specific, as this determines parameters of costs, treatment options, and rates of non-response and adverse drug reactions. Studies are also time-specific, as their conclusions may become outdated through changes in price, in management of adverse drug reactions or through the availability of new drugs.

In the context of PGx testing, the type of test applied may differ over time and between countries, and this may influence the study result. For example, a lab-developed test is likely to be less expensive than a PGx test which has undergone regulatory approval. Likewise, a multi-variant test may be less expensive than a series of single tests. For example, the PGx GeneSight test uses 44 genetic variants to guide selection of antidepressants for major depressive disorder, with some evidence that the genetic test resulted in higher response rates and was cost-saving.[28,29] Indeed, another shift in perspective may occur when PGx information is available for multiple drugs used to treat a specific condition; cost-effectiveness studies will then move from assessing a single drug to evaluating cost-effectiveness at the disease level.

Taken together, these issues imply that cost-effectiveness analyses on their own cannot answer the question of whether or not a certain strategy should be used and funded, but should be considered in conjunction with other factors such as the available resources, the number of patients who benefit from the intervention and other ethical considerations.

Warfarin provides a useful illustration of some of these issues. PGx-guided warfarin dosing was favoured by a US cost-effectiveness study but not supported by a UK study. The UK study compared warfarin (with and without PGx testing), rivaroxaban, apixaban and dabigatran, with costs and health outcomes included from the National Health Service's perspective.[22] The US study contrasted warfarin treatment without PGx testing with a strategy where all patients are tested and either receive PGx-guided doses of warfarin or an alternative drug if they have low or high warfarin sensitivity.[30] The latter analysis took the perspective of the US health-care payers. Both studies estimated costs and benefits on a lifetime horizon, measured health outcomes primarily in terms of QALYs and used *CYP2C9* and *VKORC1* for PGx testing. The UK study concluded that PGx-guided warfarin was cost-effective compared to warfarin with clinically guided dosing, but recommended the use of apixaban, which does not require PGx testing, as the most cost-effective treatment option. In contrast, the US study found that PGx-guided warfarin was cost-effective compared to clinically dosed warfarin and supported the use of PGx testing for warfarin dosing. The US study did not include apixaban or other comparator drugs, which may have influenced the conclusion reached, but many other factors differed between the studies. For example, although the price of the genetic test was twice as high in the US study, this was outweighed by differences in lifetime costs for warfarin in the United States and the United Kingdom. This example highlights the variable factors involved in performing cost-effectiveness analyses, interpreting their results and comparing such studies.

The PGx dosing algorithm for warfarin is often presented as the poster child for the achievements of PGx, because the drug is widely prescribed and implementation of this single-nucleotide polymorphism-based test could have a major impact on health care. However, only one-quarter of studies considered genetic-guided dosing for warfarin to be cost-effective, and the clinical advantage of genetic-guided dosing over standard dosing appears to be small or even non-existent.[31] Although freely available genetic testing would improve the cost-effectiveness of genotype-guided warfarin dosing, other drugs such as abacavir, where genetic testing for *HLA-B*5701* is required by the FDA, might make more convincing PGx success stories.[32]

Our study assessed the characteristics of tests in the reviewed evaluations. We noted that quoted prices for genetic tests in the United States and Canada were higher than that in other countries, although there was also a large between-study variability within these countries. However, higher prices for genetic testing in the United States and Canada did not lead to fewer conclusions in favour of PGx testing, as country was not associated with study outcome. In addition, neither the drug nor the study perspective was significantly associated with the price of testing. Genetic test costs may depend on the method used to determine genetic variants (for example, PCR or measuring enzyme activity), but the reviewed studies did not provide sufficient detail to investigate the impact of this parameter on price. A downward trend in prices for genetic testing is apparent in recent years, and this may continue as new genetic technologies become more accessible and lead to further price reductions.

We show in this study that the cost of genetic testing is an important factor in determining the cost-effectiveness of a PGx-guided treatment strategy. If there was no cost attached to genetic testing, the number of economic evaluations that found the PGx strategy cost-effective increased greatly, such that half of the reviewed studies considered it dominant over the alternative and 75% considered it cost-effective. Freely available genetic testing might be achievable in future as genomic prices fall and the perceived or actual value of genetic information increases. Once genetic tests become a mainstream clinical service, economies of scale will decrease the price of testing still further. For example, the direct to consumer testing company 23andMe offers a genome-wide genotyping service for £149 (United Kingdom, January 2017 price), which includes single-nucleotide polymorphism-based testing for 5 of the 10 drugs covered in this review.[33,34] Similarly, the cost of whole-genome sequencing has fallen every year and is now nearing US$1000.[35] Having genetic information in the electronic health record would allow PGx information to be queried for any new prescription or dosage review. A genetic test would need to be performed only once and this information, safely secured and immediately accessible, could guide treatment throughout the patient's life.

Even so, PGx-guided treatment will not be cost-effective in all situations. Even under the favourable assumption of freely available genetic testing, it could still be more expensive than the alternative strategy. This sounds counter-intuitive, but genetic testing costs may only be a small part of the costs attached to

PGx-informed treatment. Increased costs may arise where the alternative drug for test-positive patients is more expensive, and this is exacerbated whether the test has a high proportion of false-positive results. For example, patients with heart disease or stroke who are *CYP2C19* poor metabolizers may be prescribed the more expensive ticagrelor in place of clopidogrel (which is metabolized into its active form by *CYP2C19*).[25] Thus, even if genetic information is freely accessible, economic evaluations of PGx testing are still relevant and necessary.

The economic evaluation studies reviewed here show that PGx has a positive impact on health-care quality and costs. Over half of reviewed studies concluded that the PGx-informed treatment strategy is more cost-effective than the alternatives considered under present-day economics. Only one in four economic evaluations found the genetic testing option unequivocally not cost-effective. This encouraging finding, with an even bigger projected benefit under low-cost genetic typing, suggests that PGx testing has the potential to be a cost-effective or even cost-saving intervention. It therefore seems likely that PGx testing will become a core clinical service, particularly as projects such as the 100 000 Genomes Project pushes genomics to become part of health-care infrastructure and as electronic health records become increasingly effective.[36]

## CONFLICT OF INTEREST

MV is funded by a studentship from the Medical Research Council and Eli Lilly and Company Ltd. MEW is a part-time employee of Genomics plc. CML declares no conflict of interest.

## ACKNOWLEDGMENTS

This study was funded by an industrial CASE studentship to MV from the Medical Research Council with Eli Lilly and Company Ltd. This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

## REFERENCES

1 Mallal S, Phillips E, Carosi G, Molina J-M, Workman C, Tomažič J *et al.* HLA-B* 5701 screening for hypersensitivity to abacavir. *N Engl J Med* 2008; **358**: 568–579.
2 Chung W-H, Hung S-I, Hong H-S, Hsih M-S, Yang L-C, Ho H-C *et al.* Medical genetics: a marker for Stevens–Johnson syndrome. *Nature* 2004; **428**: 486–486.
3 Wadelius M, Chen LY, Eriksson N, Bumpstead S, Ghori J, Wadelius C *et al.* Association of warfarin dose with genes involved in its action and metabolism. *Hum Genet* 2007; **121**: 23–34.
4 U.S. Food and Drug Administration. Table of Pharmacogenomic Biomarkers in Drug Labeling. Available at http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm (accessed on 18 September 2015).
5 Centers for Disease Control and Prevention. Genomic Tests and Family Health History by Levels of Evidence: Public Health Genomics. Available at http://www.cdc.gov/genomics/gtesting/tier.htm (accessed on 17 February 2016).
6 Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M *et al.* The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res* 2013; **41**: 925–935.
7 Haycox A, Pirmohamed M, McLeod C, Houten R, Richards S. Through a glass darkly: economics and personalised medicine. *Pharmacoeconomics* 2014; **32**: 1055–1061.
8 Phillips KA, Van Bebber SL. A systematic review of cost-effectiveness analyses of pharmacogenomic interventions. *Pharmacogenomics* 2004; **5**: 1139–1149.
9 Vegter S, Boersma C, Rozenbaum M, Wilffert B, Navis G, Postma MJ. Pharmacoeconomic evaluations of pharmacogenetic and genomic screening programmes: a systematic review on content and adherence to guidelines. *PharmacoEconomics* 2008; **26**: 569–587.
10 Hatz MHM, Schremser K, Rogowski WH. Is individualized medicine more cost-effective? A systematic review. *PharmacoEconomics* 2014; **32**: 443–455.
11 Phillips KA, Ann Sakowski J, Trosman J, Douglas MP, Liang S-Y, Neumann P. The economic value of personalized medicine tests: what we know and what we need to know. *Genet Med* 2014; **16**: 251–257.
12 Beaulieu M, de Denus S, Lachaine J. Systematic review of pharmacoeconomic studies of pharmacogenomic tests. *Pharmacogenomics* 2010; **11**: 1573–1590.
13 Payne K, Shabaruddin FH. Cost-effectiveness analysis in pharmacogenomics. *Pharmacogenomics* 2010; **11**: 643–646.
14 Shabaruddin FH, Fleeman ND, Payne K. Economic evaluations of personalized medicine: existing challenges and current developments. *PharmacoEcon Personalized Med* 2015; **8**: 115–126.
15 Vegter S, Jansen E, Postma MJ, Boersma C. Economic evaluations of pharmacogenetic and genomic screening programs: update of the literature. *Drug Dev Res* 2010; **71**: 492–501.
16 Wong WB, Carlson JJ, Thariani R, Veenstra DL. Cost effectiveness of pharmacogenomics: a critical and systematic review. *PharmacoEconomics* 2010; **28**: 1001–1013.
17 Nixon J, Stoykova B, Christie J, Glanville J, Kleijnen J, Drummond M. NHS Economic Evaluation Database for healthcare decision makers. *BMJ* 2000; **321**: 32.
18 University of York Centre for Reviews and Dissemination. Welcome to the CRD Database. Available at http://www.crd.york.ac.uk/CRDWeb/HomePage.asp (accessed on 30 October 2015).
19 Beresniak A, Medina-Lara A, Auray JP, De Wever A, Praet JC, Tarricone R *et al.* Validation of the underlying assumptions of the quality-adjusted life-years outcome: results from the ECHOUTCOME European project. *PharmacoEconomics* 2015; **33**: 61–69.
20 McCabe C, Claxton K, Culyer A. The NICE Cost-Effectiveness Threshold. *PharmacoEconomics* 2008; **26**: 733–744.
21 Dong D, Sung C, Finkelstein EA. Cost-effectiveness of HLA-B*1502 genotyping in adult patients with newly diagnosed epilepsy in Singapore. *Neurology* 2012; **79**: 1259–1267.
22 Pink J, Pirmohamed M, Lane S, Hughes DA. Cost-effectiveness of pharmacogenetics-guided warfarin therapy vs. alternative anticoagulation in atrial fibrillation. *Clin Pharmacol Ther* 2014; **95**: 199–207.
23 Nshimyumukiza L, Duplantie J, Gagnon M, Douville X, Fournier D, Lindsay C *et al.* Dabigatran versus warfarin under standard or pharmacogenetic-guided management for the prevention of stroke and systemic thromboembolism in patients with atrial fibrillation: a cost/utility analysis using an analytic decision model. *Thromb J* 2013; **11**: 14.
24 Patrick AR, Avorn J, Choudhry NK. Cost-effectiveness of genotype-guided warfarin dosing for patients with atrial fibrillation. *Circ Cardiovasc Qual Outcomes* 2009; **2**: 429–436.
25 Sorich MJ, Horowitz JD, Sorich W, Wiese MD, Pekarsky B, Karnon JD. Cost-effectiveness of using CYP2C19 genotype to guide selection of clopidogrel or ticagrelor in Australia. *Pharmacogenomics* 2013; **14**: 2013–2021.
26 Lala A, Berger JS, Sharma G, Hochman JS, Scott Braithwaite R, Ladapo JA. Genetic testing in patients with acute coronary syndrome undergoing percutaneous coronary intervention: a cost-effectiveness analysis. *J Thromb Haemost* 2013; **11**: 81–91.
27 Wiley Online Library, John Wiley & Sons Ltd. HEED: Health Economic Evaluations Database. Available at: http://onlinelibrary.wiley.com/book/10.1002/9780470510933 (accessed on 28 January 2016).
28 Altar CA, Carhart J, Allen JD, Hall-Flavin D, Winner J, Dechairo B. Clinical utility of combinatorial pharmacogenomics-guided antidepressant therapy: evidence from three clinical studies. *Mol Neuropsychiatry* 2015; **1**: 145–155.
29 Hornberger J, Li Q, Quinn B. Cost-effectiveness of combinatorial pharmacogenomic testing for treatment-resistant major depressive disorder patients. *Am J Manag Care* 2014; **21**: e357–e365.
30 You JH. Pharmacogenetic-guided selection of warfarin versus novel oral anticoagulants for stroke prevention in patients with atrial fibrillation: a cost-effectiveness analysis. *Pharmacogenet Genomics* 2014; **24**: 6–14.
31 Stergiopoulos K, Brown DL. Genotype-guided vs clinical dosing of warfarin and its analogues: meta-analysis of randomized clinical trials. *JAMA Intern Med* 2014; **174**: 1330–1338.
32 U.S. Food and Drug Administration. Ziagen Prescribing Information. Available at http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/020977s030,020978s034lbl.pdf (accessed on 2 February 2016).
33 23andMe Inc. 23andMe: Welcome. Available at https://www.23andme.com/en-gb (accessed on 1 February 2016).
34 Lu M, Lewis CM, Traylor M. Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. *bioRxiv* 2017; https://doi.org/10.1101/098541 (accessed 6 Feb 2017).
35 Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at www.genome.gov/sequencingcostsdata (accessed on 27 January 2016).
36 Genomics England: Genomics England is delivering the 100,000 Genomes Project. Available at http://www.genomicsengland.co.uk (accessed on 31 March 2016).
37 Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 2012; **92**: 414–417.

38 Hughes DA, Vilar FJ, Ward CC, Alfirevic A, Park BK, Pirmohamed M. Cost-effectiveness analysis of HLA B*5701 genotyping in preventing abacavir hypersensitivity. *Pharmacogenetics* 2004; **14**: 335–342.
39 Schackman BR, Scott CA, Walensky RP, Losina E, Freedberg KA, Sax PE. The cost-effectiveness of HLA-B*5701 genetic screening to guide initial antiretroviral therapy for HIV. *AIDS* 2008; **22**: 2025–2033.
40 Kauf TL, Farkouh RA, Earnshaw SR, Watson ME, Maroudas P, Chambers MG. Economic efficiency of genetic screening to inform the use of abacavir sulfate in the treatment of HIV. *Pharmacoeconomics* 2010; **28**: 1025–1039.
41 Nieves Calatrava D, Calle-Martin Ode L, Iribarren-Loyarte JA, Rivero-Roman A, Garcia-Bujalance L, Perez-Escolano I *et al.* Cost-effectiveness analysis of HLA-B*5701 typing in the prevention of hypersensitivity to abacavir in HIV+ patients in Spain. *Enferm Infecc Microbiol Clin* 2010; **28**: 590–595.
42 Kapoor R, Martinez-Vega R, Dong D, Tan SY, Leo YS, Lee CC *et al.* Reducing hypersensitivity reactions with HLA-B*5701 genotyping before abacavir prescription: clinically useful but is it cost-effective in Singapore? *Pharmacogenet Genomics* 2015; **25**: 60–72.
43 Tavadia SM, Mydlarski PR, Reis MD, Mittmann N, Pinkerton PH, Shear N *et al.* Screening for azathioprine toxicity: a pharmacoeconomic analysis based on a target case. *J Am Acad Dermatol* 2000; **42**: 628–632.
44 Marra CA, Esdaile JM, Anis AH. Practical pharmacogenetics: the cost effectiveness of screening for thiopurine s-methyltransferase polymorphisms in patients with rheumatological conditions treated with azathioprine. *J Rheumatol* 2002; **29**: 2507–2512.
45 Oh KT, Anis AH, Bae SC. Pharmacoeconomic analysis of thiopurine methyltransferase polymorphism screening by polymerase chain reaction for treatment with azathioprine in Korea. *Rheumatology (Oxford)* 2004; **43**: 156–163.
46 Winter J, Walker A, Shapiro D, Gaffney D, Spooner RJ, Mills PR. Cost-effectiveness of thiopurine methyltransferase genotype screening in patients about to commence azathioprine therapy for treatment of inflammatory bowel disease. *Aliment Pharmacol Ther* 2004; **20**: 593–599.
47 Dubinsky MC, Reyes E, Ofman J, Chiou C-F, Wade S, Sandborn WJ. A cost-effectiveness analysis of alternative disease management strategies in patients with Crohn's disease treated with azathioprine or 6-mercaptopurine. *Am J Gastroenterol* 2005; **100**: 2239–2247.
48 Priest VL, Begg EJ, Gardiner SJ, Frampton CM, Gearry RB, Barclay ML *et al.* Pharmacoeconomic analyses of azathioprine, methotrexate and prospective pharmacogenetic testing for the management of inflammatory bowel disease. *Pharmacoeconomics* 2006; **24**: 767–781.
49 Van Den Akker-Van Marle ME, Gurwitz D, Detmar SB, Enzing CM, Hopkins MM, Gutierrez De Mesa E *et al.* Cost-effectiveness of pharmacogenomics in clinical practice: a case study of thiopurine methyltransferase genotyping in acute lymphoblastic leukemia in Europe. *Pharmacogenomics* 2006; **7**: 783–792.
50 Hagaman JT, Kinder BW, Eckman MH. Thiopurine S- methyltransferase [corrected] testing in idiopathic pulmonary fibrosis: a pharmacogenetic cost-effectiveness analysis. *Lung* 2010; **188**: 125–132.
51 Thompson AJ, Newman WG, Elliott RA, Roberts SA, Tricker K, Payne K. The cost-effectiveness of a pharmacogenetic test: a trial-based evaluation of TPMT genotyping for azathioprine. *Value Health* 2014; **17**: 22–33.
52 Rattanavipapong W, Koopitakkajorn T, Praditsitthikorn N, Mahasirimongkol S, Teerawattananon Y. Economic evaluation of HLA-B*15:02 screening for carbamazepine-induced severe adverse drug reactions in Thailand. *Epilepsia* 2013; **54**: 1628–1638.
53 Tiamkao S, Jitpimolmard J, Sawanyawisuth K, Jitpimolmard S. Cost minimization of HLA-B*1502 screening before prescribing carbamazepine in Thailand. *Int J Clin Pharm* 2013; **35**: 608–612.
54 Plumpton CO, Yip VL, Alfirevic A, Marson AG, Pirmohamed M, Hughes DA. Cost-effectiveness of screening for HLA-A*31:01 prior to initiation of carbamazepine in epilepsy. *Epilepsia* 2015; **56**: 556–563.
55 McMahon FJ, Buervenich S, Charney D, Lipsky R, Rush AJ, Wilson AF *et al.* Variation in the gene encoding the serotonin 2A receptor is associated with outcome of antidepressant treatment. *Am J Hum Genet* 2006; **78**: 804–814.
56 Serretti A, Kato M, De Ronchi D, Kinoshita T. Meta-analysis of serotonin transporter gene promoter polymorphism (5-HTTLPR) association with selective serotonin reuptake inhibitor efficacy in depressed patients. *Mol Psychiatry* 2007; **12**: 247–257.
57 Perlis RH, Patrick A, Smoller JW, Wang PS. When is Pharmacogenetic Testing for Antidepressant Response Ready for the Clinic[quest] A Cost-effectiveness Analysis Based on Data from the STAR[ast]D Study. *Neuropsychopharmacology* 2009; **34**: 2227–2236.
58 Serretti A, Olgiati P, Bajo E, Bigelli M, De Ronchi D. A model to incorporate genetic testing (5-HTTLPR) in pharmacological treatment of major depressive disorders. *World J Biol Psychiatry* 2011; **12**: 501–515.
59 Olgiati P, Bajo E, Bigelli M, De Ronchi D, Serretti A. Should pharmacogenetics be incorporated in major depression treatment? Economic evaluation in high- and middle-income European countries. *Prog Neuropsychopharmacol Biol Psychiatry* 2012; **36**: 147–154.
60 Crespin DJ, Federspiel JJ, Biddle AK, Jonas DE, Rossi JS. Ticagrelor versus genotype-driven antiplatelet therapy for secondary prevention after acute coronary syndrome: a cost-effectiveness analysis. *Value Health* 2011; **14**: 483–491.
61 Panattoni L, Brown PM, Te Ao B, Webster M, Gladding P. The cost effectiveness of genetic testing for CYP2C19 variants to guide thienopyridine treatment in patients with acute coronary syndromes: a New Zealand evaluation. *Pharmacoeconomics* 2012; **30**: 1067–1084.
62 Reese ES, Daniel Mullins C, Beitelshees AL, Onukwugha E. Cost-effectiveness of cytochrome P450 2C19 genotype screening for selection of antiplatelet therapy with clopidogrel or prasugrel. *Pharmacotherapy* 2012; **32**: 323–332.
63 Kazi DS, Garber AM, Shah RU, Dudley RA, Mell MW, Rhee C *et al.* Cost-effectiveness of genotype-guided and dual antiplatelet therapies in acute coronary syndrome. *Ann Intern Med* 2014; **160**: 221–232.
64 Arranz MJ, Munro J, Birkett J, Bolonna A, Mancama D, Sodhi M *et al.* Pharmacogenetic prediction of clozapine response. *Lancet* 2000; **355**: 1615–1616.
65 Perlis RH, Ganz DA, Avorn J, Schneeweiss S, Glynn RJ, Smoller JW *et al.* Pharmacogenetic testing in the clinical management of schizophrenia: a decision-analytic model. *J Clin Psychopharmacol* 2005; **25**: 427–434.
66 Obradovic M, Mrhar A, Kos M. Cost-effectiveness of UGT1A1 genotyping in second-line, high-dose, once every 3 weeks irinotecan monotherapy treatment of colorectal cancer. *Pharmacogenomics* 2008; **9**: 539–549.
67 Gold HT, Hall MJ, Blinder V, Schackman BR. Cost effectiveness of pharmacogenetic testing for uridine diphosphate glucuronosyltransferase 1A1 before irinotecan administration for metastatic colorectal cancer. *Cancer* 2009; **115**: 3858–3867.
68 Pichereau S, Le Louarn A, Lecomte T, Blasco H, Le Guellec C, Bourgoin H. Cost-effectiveness of UGT1A1*28 genotyping in preventing severe neutropenia following FOLFIRI therapy in colorectal cancer. *J Pharm Pharm Sci* 2010; **13**: 615–625.
69 Donnan JR, Ungar WJ, Mathews M, Hancock-Howard RL, Rahman P. A cost effectiveness analysis of thiopurine methyltransferase testing for guiding 6-mercaptopurine dosing in children with acute lymphoblastic leukemia. *Pediatr Blood Cancer* 2011; **57**: 231–239.
70 You JH, Chan FW, Wong RS, Cheng G. The potential clinical and economic outcomes of pharmacogenetics-oriented management of warfarin therapy—a decision analysis. *Thromb Haemost* 2004; **92**: 590–597.
71 Eckman MH, Rosand J, Greenberg SM, Gage BF. Cost-effectiveness of using pharmacogenetic information in warfarin dosing for patients with nonvalvular atrial fibrillation. *Ann Intern Med* 2009; **150**: 73–83.
72 Leey JA, McCabe S, Koch JA, Miles TP. Cost-effectiveness of genotype-guided warfarin therapy for anticoagulation in elderly patients with atrial fibrillation. *Am J Geriatr Pharmacother* 2009; **7**: 197–203.
73 You JH, Tsui KK, Wong RS, Cheng G. Potential clinical and economic outcomes of CYP2C9 and VKORC1 genotype-guided dosing in patients starting warfarin therapy. *Clin Pharmacol Ther* 2009; **86**: 540–547.
74 Meckley L, Gudgeon J, Anderson J, Williams M, Veenstra D. A policy model to evaluate the benefits, risks and costs of warfarin pharmacogenomic testing. *PharmacoEconomics* 2010; **28**: 61–74.
75 You JHS, Tsui KKN, Wong RSM, Cheng G. Cost-effectiveness of dabigatran versus genotype-guided management of warfarin therapy for stroke prevention in patients with atrial fibrillation. *PLoS ONE* 2012; **7**: e39640.
76 Chong HY, Saokaew S, Dumrongprat K, Permsuwan U, Wu DB, Sritara P *et al.* Cost-effectiveness analysis of pharmacogenetic-guided warfarin dosing in Thailand. *Thromb Res* 2014; **134**: 1278–1284.
77 Mitropoulou C, Fragoulakis V, Bozina N, Vozikis A, Supe S, Bozina T *et al.* Economic evaluation of pharmacogenomic-guided warfarin treatment for elderly Croatian atrial fibrillation patients with ischemic stroke. *Pharmacogenomics* 2015; **16**: 137–148.



Supplementary Information accompanies the paper on the The Pharmacogenomics Journal website (http://www.nature.com/tpj)

## Supplementary material

Supplementary table 1a. Column headings for supplementary table 1b.

| Column name | Column explanation |
|---|---|
| Authors | First author and reference to citation |
| Year of publication | Year of publication |
| Country | Country where economic evaluation was performed |
| Perspective | Perspective used when defining costs and effects to include in evaluation |
| Drug | Generic name of drug for which a PGx test is studied |
| Gene | Gene(s) used in PGx test |
| Type of study | The type of economic evaluation: cost-effectiveness analysis (CEA), cost-utility analysis (CUA) or cost-benefit analysis (CBA) |
| Outcome measure | The measure used to quantify the effect of treatment |
| Sensitivity analysis | Extent of sensitivity analysis performed in the study |
| Price of genetic test quoted in study | Price of the genetic test as reported be study |
| Price of genetic test in 2014 US$ | Price of genetic test corrected for inflation and converted to 2014US$ |
| Result | Result reported by study. Incremental cost-effectiveness ratio (ICER) if applicable. |
| Estimated result | Our estimated result if genetic information was freely available. Estimated by taking the cost of the PGx test as reported out of the total costs reported by the study. |
| Conclusion for PGx strategy | Conclusion reached by study. Classified as: Favourable and cost saving (dominant), favourable (cost-effective, not cost-saving), not favourable (not cost-effective) or undetermined (no clear conclusion). |
| Estimated conclusion for PGx strategy | Our estimated conclusion if genetic information was freely available. Based on the estimated result. |

Supplementary table 1b. Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Year of publication | Country | Perspective |
|---|---|---|---|
| Hughes et al.[38] | 2004 | UK | National Health Service |
| Schackman et al.[39] | 2008 | US | Not stated |
| Kauf et al.[40] | 2010 | US | US healthcare system |
| Nieves Calatrava et al.[41] | 2010 | Spain | National Health System |
| Kapoor et al.[42] | 2015 | Singapore | Not stated |
| Tavadia et al.[43] | 2000 | Canada | Not stated |
| Marra et al.[44] | 2002 | Canada | Third party payer |
| Oh et al.[45] | 2004 | Korea | Society |
| Winter et al.[46] | 2004 | Scotland | Not stated |
| Dubinsky et al.[47] | 2005 | US | Third party payer |
| Priest et al.[48] | 2006 | New Zealand | New Zealand government and patients with inflammatory bowel disease |
| Van Den Akker-Van Marle et al.[49] | 2006 | Germany, Ireland, The Netherlands and UK | Society |
| Hagaman et al.[50] | 2010 | US | Not stated |
| Thompson et al.[51] | 2014 | UK | UK health service |
| Dong et al.[21] | 2012 | Singapore | Not stated |
| Rattanavipapong et al.[52] | 2013 | Thailand | Society |
| Tiamkao et al.[53] | 2013 | Thailand | Not stated |
| Plumpton et al.[54] | 2015 | UK | National Health Service |
| Perlis et al.[57] | 2009 | US | Society |
| Serretti et al.[58] | 2011 | Italy | Italian National Health System |
| Olgiati et al.[59] | 2012 | Europe | Society |
| Crespin et al.[60] | 2011 | US | Medicare |
| Panattoni et al.[61] | 2012 | New Zealand | New Zealand healthcare system |
| Reese et al.[62] | 2012 | US | Private payer |
| Lala et al.[26] | 2013 | US | Payer |
| Sorich et al.[25] | 2013 | Australia | Australian healthcare system |
| Kazi et al.[63] | 2014 | US | Society |
| Perlis et al.[65] | 2005 | US | Society |
| Obradovic et al.[66] | 2008 | US | US healthcare payers |
| Gold et al.[67] | 2009 | US | Medicare payer |
| Pichereau et al.[68] | 2010 | France | Hospital |
| Donnan et al.[69] | 2011 | Canada | Healthcare system |
| You et al.[70] | 2004 | Not clearly stated | Healthcare provider |
| Eckman et al.[71] | 2009 | US | Society |
| Leey et al.[72] | 2009 | US | third-party payer |
| Patrick et al.[24] | 2009 | US | Society |
| You et al.[73] | 2009 | Not clearly stated | Healthcare provider |
| Meckley et al.[74] | 2010 | US | US third party payer |
| You et al.[75] | 2012 | Not clearly stated | Healthcare payers |
| Nshimyumukiza et al.[23] | 2013 | Canada | Public Health system |
| Chong et al.[76] | 2014 | Thailand | Healthcare system and society |
| Pink et al.[22] | 2014 | UK | National Health Service |
| You[30] | 2014 | Not clearly stated | US healthcare payers |
| Mitropoulou et al.[77] | 2015 | Croatia | Sickness Fund perspective, healthcare payers |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Drug | Gene | Type of study |
|---|---|---|---|
| Hughes et al.[38] | abacavir | *HLA* | CEA |
| Schackman et al.[39] | abacavir | *HLA* | CUA |
| Kauf et al.[40] | abacavir | *HLA* | CUA |
| Nieves Calatrava et al.[41] | abacavir | *HLA* | CEA |
| Kapoor et al.[42] | abacavir | *HLA* | CUA |
| Tavadia et al.[43] | azathioprine | *TPMT* | CBA |
| Marra et al.[44] | azathioprine | *TPMT* | CEA |
| Oh et al.[45] | azathioprine | *TPMT* | CEA |
| Winter et al.[46] | azathioprine | *TPMT* | CEA |
| Dubinsky et al.[47] | azathioprine | *TPMT* | CEA |
| Priest et al.[48] | azathioprine | *TPMT* | CUA |
| Van Den Akker-Van Marle et al.[49] | azathioprine | *TPMT* | CEA |
| Hagaman et al.[50] | azathioprine | *TPMT* | CUA |
| Thompson et al.[51] | azathioprine | *TPMT* | CUA |
| Dong et al.[21] | carbamazepine and phenytoin | *HLA* | CUA |
| Rattanavipapong et al.[52] | carbamazepine | *HLA* | CUA |
| Tiamkao et al.[53] | carbamazepine | *HLA* | CBA |
| Plumpton et al.[54] | carbamazepine | *HLA* | CUA |
| Perlis et al.[57] | citalopram | *HTR2A* | CUA |
| Serretti et al.[58] | citalopram | *5-HTTLPR* | CUA |
| Olgiati et al.[59] | citalopram | *5-HTTLPR* | CUA |
| Crespin et al.[60] | clopidogrel | *CYP2C19* | CUA |
| Panattoni et al.[61] | clopidogrel | *CYP2C19* | CUA |
| Reese et al.[62] | clopidogrel | *CYP2C19* | CEA |
| Lala et al.[26] | clopidogrel | *CYP2C19* | CUA |
| Sorich et al.[25] | clopidogrel | *CYP2C19* | CUA |
| Kazi et al.[63] | clopidogrel | *CYP2C19* | CUA |
| Perlis et al.[65] | clozapine | *H2, 5-HTTLPR, 5-HT$_{2A}$, 5-HT$_{2C}$* | CUA |
| Obradovic et al.[66] | irinotecan | *UGT1A1* | CEA |
| Gold et al.[67] | irinotecan | *UGT1A1* | CUA |
| Pichereau et al.[68] | irinotecan | *UGT1A1* | CEA |
| Donnan et al.[69] | mercaptopurine | *TPMT* | CEA |
| You et al.[70] | warfarin | *CYP2C9* | CEA |
| Eckman et al.[71] | warfarin | *CYP2C9, VKORC1* | CUA |
| Leey et al.[72] | warfarin | *CYP2C9, VKORC1* | CUA |
| Patrick et al.[24] | warfarin | *CYP2C9, VKORC1* | CUA |
| You et al.[73] | warfarin | *CYP2C9, VKORC1* | CUA |
| Meckley et al.[74] | warfarin | *CYP2C9, VKORC1* | CUA |
| You et al.[75] | warfarin | *CYP2C9, VKORC1* | CUA |
| Nshimyumukiza et al.[23] | warfarin | *CYP2C9, VKORC1* | CUA |
| Chong et al.[76] | warfarin | *CYP2C9, VKORC1* | CUA |
| Pink et al.[22] | warfarin | *CYP2C9, VKORC1* | CUA |
| You[30] | warfarin | *CYP2C9, VKORC1* | CUA |
| Mitropoulou et al.[77] | warfarin | *CYP2C9, VKORC1* | CUA |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Outcome measure | Sensitivity analysis |
|---|---|---|
| Hughes et al.[38] | adverse reactions avoided | Univariate |
| Schackman et al.[39] | QALY | Univariate and multivariate |
| Kauf et al.[40] | QALY | Univariate |
| Nieves Calatrava et al.[41] | cost per adverse reaction avoided | Univariate and multivariate |
| Kapoor et al.[42] | QALY | Univariate and multivariate |
| Tavadia et al.[43] | Canadian dollars | Univariate |
| Marra et al.[44] | number needed to treat to avoid one adverse event | Univariate |
| Oh et al.[45] | probability of not dropping out due to serious adverse event | Univariate and multivariate |
| Winter et al.[46] | life years saved | None reported |
| Dubinsky et al.[47] | time to response | Univariate and multivariate |
| Priest et al.[48] | QALY | Univariate |
| Van Den Akker-Van Marle et al.[49] | life years gained | Univariate and multivariate |
| Hagaman et al.[50] | QALY | Univariate and multivariate |
| Thompson et al.[51] | QALY | Univariate |
| Dong et al.[21] | QALY | Univariate and multivariate |
| Rattanavipapong et al.[52] | QALY | Multivariate |
| Tiamkao et al.[53] | Thai baht | None reported |
| Plumpton et al.[54] | QALY | Univariate and multivariate |
| Perlis et al.[57] | QALY | Univariate and multivariate |
| Serretti et al.[58] | QALWeek | Univariate and multivariate |
| Olgiati et al.[59] | QALWeek | Univariate and multivariate |
| Crespin et al.[60] | QALY | Univariate and multivariate |
| Panattoni et al.[61] | QALY | Multivariate |
| Reese et al.[62] | adverse events avoided | Multivariate |
| Lala et al.[26] | QALY | Univariate and multivariate |
| Sorich et al.[25] | QALY | Univariate and multivariate |
| Kazi et al.[63] | QALY | Univariate and multivariate |
| Perlis et al.[65] | QALY | Univariate |
| Obradovic et al.[66] | life years gained | Multivariate |
| Gold et al.[67] | QALY | Univariate and multivariate |
| Pichereau et al.[68] | number of neutropenias avoided | Univariate and multivariate |
| Donnan et al.[69] | life months survived | Univariate and multivariate |
| You et al.[70] | major bleedings averted | Univariate and multivariate |
| Eckman et al.[71] | QALY | Univariate and multivariate |
| Leey et al.[72] | QALY | Multivariate |
| Patrick et al.[24] | QALY | Univariate and multivariate |
| You et al.[73] | QALY | Univariate and multivariate |
| Meckley et al.[74] | QALY | Univariate and multivariate |
| You et al.[75] | QALY | Univariate and multivariate |
| Nshimyumukiza et al.[23] | QALY | Univariate and multivariate |
| Chong et al.[76] | QALY | Univariate and multivariate |
| Pink et al.[22] | QALY | Univariate and multivariate |
| You[30] | QALY | Univariate and multivariate |
| Mitropoulou et al.[77] | QALY | Multivariate |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Price of genetic test quoted in study | Price of genetic test in 2014 US$ |
|---|---|---|
| Hughes et al.[38] | €43.40 | US$77.18 |
| Schackman et al.[39] | US$68.00 | US$79.85 |
| Kauf et al.[40] | US$87.92 | US$100.39 |
| Nieves Calatrava et al.[41] | €55.00 | US$80.19 |
| Kapoor et al.[42] | US$277.00 | US$275.73 |
| Tavadia et al.[43] | C$100.00 | US$118.59 |
| Marra et al.[44] | C$100.00 | US$112.95 |
| Oh et al.[45] | ₩60000.00 | US$79.58 |
| Winter et al.[46] | £30.00 | US$67.79 |
| Dubinsky et al.[47] | US$510.06 | US$639.21 |
| Priest et al.[48] | NZ$120.00 | US$128.08 |
| Van Den Akker-Van Marle et al.[49] | €150.00 | US$240.55 |
| Hagaman et al.[50] | US$300.00 | US$342.54 |
| Thompson et al.[51] | £20.00 | US$39.48 |
| Dong et al.[21] | US$270.00 | US$293.14 |
| Rattanavipapong et al.[52] | ฿1000.00 | US$33.03 |
| Tiamkao et al.[53] | ฿3000.00 | US$94.12 |
| Plumpton et al.[54] | £54.26 | US$102.39 |
| Perlis et al.[57] | US$500.00 | US$587.15 |
| Serretti et al.[58] | 2010 Int$233.80 | US$253.84 |
| Olgiati et al.[59] | 2009 Int$200.00 | US$220.70 |
| Crespin et al.[60] | US$200.00 | US$220.70 |
| Panattoni et al.[61] | NZ$175.00 | US$161.17 |
| Reese et al.[62] | US$310.00 | US$326.24 |
| Lala et al.[26] | US$500.00 | US$542.85 |
| Sorich et al.[25] | AUS$46.55 | US$44.90 |
| Kazi et al.[63] | US$235.00 | US$247.31 |
| Perlis et al.[65] | US$500.00 | US$710.50 |
| Obradovic et al.[66] | US$375.00 | US$440.36 |
| Gold et al.[67] | US$102.83 | US$117.41 |
| Pichereau et al.[68] | €71.00 | US$106.01 |
| Donnan et al.[69] | C$459.63 | US$452.73 |
| You et al.[70] | US$100.00 | US$125.32 |
| Eckman et al.[71] | US$400.00 | US$456.72 |
| Leey et al.[72] | US$250.00 | US$321.65 |
| Patrick et al.[24] | US$575.00 | US$656.54 |
| You et al.[73] | US$200.00 | US$219.90 |
| Meckley et al.[74] | US$175.00 | US$199.82 |
| You et al.[75] | US$72.00 | US$74.24 |
| Nshimyumukiza et al.[23] | C$615.00 | US$599.53 |
| Chong et al.[76] | ฿3000.00 | US$94.12 |
| Pink et al.[22] | £20.00 | US$35.87 |
| You[30] | US$75.00 | US$76.22 |
| Mitropoulou et al.[77] | €140.25 | US$189.66 |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Result |
|---|---|
| Hughes et al.[38] | Dominant - €22 811/adverse reaction avoided |
| Schackman et al.[39] | US$36 700/QALY |
| Kauf et al.[40] | Dominant |
| Nieves Calatrava et al.[41] | €630/adverse reaction avoided |
| Kapoor et al.[42] | US$44 649 - US$926 938/QALY |
| Tavadia et al.[43] | US$1.30 - US$233.90 saved per patient tested |
| Marra et al.[44] | Dominant |
| Oh et al.[45] | Dominant |
| Winter et al.[46] | £347/life year saved (30 year old) - £817/life year saved (60 year old) |
| Dubinsky et al.[47] | Dominant |
| Priest et al.[48] | Dominated |
| Van Den Akker-Van Marle et al.[49] | €2 100/life year gained |
| Hagaman et al.[50] | US$29 662/QALY |
| Thompson et al.[51] | £256.89 expected incremental net benefit |
| Dong et al.[21] | US$7 930 - US$136 630/QALY |
| Rattanavipapong et al.[52] | ฿130 000/QALY for patients with neuropathic pain and ฿222 000/QALY for patients with epilepsy |
| Tiamkao et al.[53] | ฿98 549.94 saved per 100 tested patients |
| Plumpton et al.[54] | £12 808/QALY |
| Perlis et al.[57] | US$93 520/QALY |
| Serretti et al.[58] | Int$2 890/QALW for 1 episode and Int$1 392/QALW for 2 recurrent episodes |
| Olgiati et al.[59] | Int$1 147 - Int$1 185/QALW |
| Crespin et al.[60] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Panattoni et al.[61] | NZ$24 617/QALY |
| Reese et al.[62] | Dominant - US$2 300/adverse event avoided |
| Lala et al.[26] | Dominant |
| Sorich et al.[25] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Kazi et al.[63] | US$30 200/QALY |
| Perlis et al.[65] | US$47 705/QALY |
| Obradovic et al.[66] | Dominant - US$6 818 203/life year gained |
| Gold et al.[67] | Dominant |
| Pichereau et al.[68] | €942.80 - €1 090.10 per neutropenia avoided |
| Donnan et al.[69] | Dominated |
| You et al.[70] | US$5 778/major bleeding averted |
| Eckman et al.[71] | US$170 000/QALY |
| Leey et al.[72] | Any reduction in major bleeding would offset the higher costs of PGx testing. |
| Patrick et al.[24] | ICER <US$50,000/QALY if genotyping increases time spent in therapeutic INR range by 9% |
| You et al.[73] | US$347 059/QALY |
| Meckley et al.[74] | US$60 725/QALY |
| You et al.[75] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Nshimyumukiza et al.[23] | Dominated |
| Chong et al.[76] | ฿1 473 852/QALY from societal perspective and ฿1 477 042/QALY from healthcare system perspective |
| Pink et al.[22] | Alternative strategy is cost-effective compared to PGx guided treatment |
| You[30] | US$2 843/QALY |
| Mitropoulou et al.[77] | €31 225/QALY |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Estimated result |
|---|---|
| Hughes et al.[38] | Dominant - €19 811/adverse reaction avoided |
| Schackman et al.[39] | US$12 600/QALY |
| Kauf et al.[40] | Dominant |
| Nieves Calatrava et al.[41] | Dominant |
| Kapoor et al.[42] | Dominant - US$164 127/QALY |
| Tavadia et al.[43] | US$101.30 - US$333.90 saved per patient tested |
| Marra et al.[44] | Dominant |
| Oh et al.[45] | Dominant |
| Winter et al.[46] | Dominant |
| Dubinsky et al.[47] | Dominant |
| Priest et al.[48] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Van Den Akker-Van Marle et al.[49] | Dominant |
| Hagaman et al.[50] | Not possible to estimate |
| Thompson et al.[51] | Not possible to estimate |
| Dong et al.[21] | US$4 416 - US$34 221/QALY |
| Rattanavipapong et al.[52] | ฿94 970/QALY for patients with neuropathic pain and ฿190 104/QALY for patients with epilepsy |
| Tiamkao et al.[53] | ฿398 549.94 saved per 100 tested patients |
| Plumpton et al.[54] | £10 502/QALY |
| Perlis et al.[57] | US$1 019/QALY |
| Serretti et al.[58] | Dominant |
| Olgiati et al.[59] | Dominant |
| Crespin et al.[60] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Panattoni et al.[61] | NZ$15 737/QALY |
| Reese et al.[62] | Not possible to estimate |
| Lala et al.[26] | Dominant |
| Sorich et al.[25] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Kazi et al.[63] | US$25 931/QALY |
| Perlis et al.[65] | Not possible to estimate |
| Obradovic et al.[66] | Dominant |
| Gold et al.[67] | Dominant |
| Pichereau et al.[68] | €162.64 - €309.89 per neutropenia avoided per 1000 patients |
| Donnan et al.[69] | Dominant |
| You et al.[70] | Not possible to estimate |
| Eckman et al.[71] | Dominant |
| Leey et al.[72] | Not possible to estimate |
| Patrick et al.[24] | Not possible to estimate |
| You et al.[73] | Dominant |
| Meckley et al.[74] | Dominant |
| You et al.[75] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Nshimyumukiza et al.[23] | Alternative strategy is cost-effective compared to PGx guided treatment |
| Chong et al.[76] | Dominant |
| Pink et al.[22] | Alternative strategy is cost-effective compared to PGx guided treatment |
| You[30] | US$2 450/QALY |
| Mitropoulou et al.[77] | €17 512/QALY |

Supplementary table 1b (continued). Overview of reviewed economic evaluations of PGx guided strategies.

| Authors | Conclusion for PGx strategy | Estimated conclusion for PGx strategy |
|---|---|---|
| Hughes et al.[38] | Favourable, cost-saving | Favourable, cost-saving |
| Schackman et al.[39] | Favourable | Favourable |
| Kauf et al.[40] | Favourable, cost-saving | Favourable, cost-saving |
| Nieves Calatrava et al.[41] | Favourable | Favourable, cost-saving |
| Kapoor et al.[42] | Not favourable | Undetermined |
| Tavadia et al.[43] | Favourable, cost-saving | Favourable, cost-saving |
| Marra et al.[44] | Favourable, cost-saving | Favourable, cost-saving |
| Oh et al.[45] | Favourable, cost-saving | Favourable, cost-saving |
| Winter et al.[46] | Favourable | Favourable, cost-saving |
| Dubinsky et al.[47] | Favourable, cost-saving | Favourable, cost-saving |
| Priest et al.[48] | Not favourable | Not favourable |
| Van Den Akker-Van Marle et al.[49] | Favourable | Favourable, cost-saving |
| Hagaman et al.[50] | Favourable | Favourable |
| Thompson et al.[51] | Undetermined | Not possible to estimate |
| Dong et al.[21] | Favourable | Favourable |
| Rattanavipapong et al.[52] | Undetermined | Undetermined |
| Tiamkao et al.[53] | Favourable, cost-saving | Favourable, cost-saving |
| Plumpton et al.[54] | Favourable | Favourable |
| Perlis et al.[57] | Not favourable | Favourable |
| Serretti et al.[58] | Undetermined | Favourable, cost-saving |
| Olgiati et al.[59] | Undetermined | Favourable, cost-saving |
| Crespin et al.[60] | Not favourable | Not favourable |
| Panattoni et al.[61] | Favourable | Favourable |
| Reese et al.[62] | Favourable, cost-saving | Favourable, cost-saving |
| Lala et al.[26] | Favourable, cost-saving | Favourable, cost-saving |
| Sorich et al.[25] | Not favourable | Not favourable |
| Kazi et al.[63] | Favourable | Favourable |
| Perlis et al.[65] | Favourable | Favourable |
| Obradovic et al.[66] | Favourable, cost-saving | Favourable, cost-saving |
| Gold et al.[67] | Favourable, cost-saving | Favourable, cost-saving |
| Pichereau et al.[68] | Favourable | Favourable |
| Donnan et al.[69] | Not favourable | Favourable, cost-saving |
| You et al.[70] | Undetermined | Not possible to estimate |
| Eckman et al.[71] | Not favourable | Favourable, cost-saving |
| Leey et al.[72] | Favourable, cost-saving | Favourable, cost-saving |
| Patrick et al.[24] | Undetermined | Not possible to estimate |
| You et al.[73] | Not favourable | Favourable, cost-saving |
| Meckley et al.[74] | Undetermined | Favourable, cost-saving |
| You et al.[75] | Undetermined | Undetermined |
| Nshimyumukiza et al.[23] | Not favourable | Not favourable |
| Chong et al.[76] | Not favourable | Favourable, cost-saving |
| Pink et al.[22] | Not favourable | Not favourable |
| You[30] | Favourable | Favourable |
| Mitropoulou et al.[77] | Favourable | Favourable |

# Part 2: Machine learning prediction algorithms applied to genetic and gene expression studies

In the second part of this thesis, we used machine learning methods to perform multivariable analyses of genetic and gene expression studies. Various statistical approaches were applied to genetic and gene expression datasets and we start by giving an overview of the statistical methods used. Next, we expand on three studies that were undertaken using machine learning algorithms: the classification of schizophrenia cases from controls using brain gene expression scores, a PGx analysis of an anti-diabetic clinical trial and a PGx analysis of an antidepressant clinical trial.

# 5. Statistical methods

A range of statistical approaches were used for the analysis of genetic and RNA-seq gene expression datasets. This methods chapter starts with an overview of the notation used throughout this thesis. Next, the traditional statistical models applied in the thesis are described. The key concepts of machine learning are defined and then a detailed introduction to the specific machine learning methods and deep learning algorithms used in this thesis are given.

## 5.1. Notation

Unless otherwise stated, the following notation will be used to describe statistical methods. For cross-sectional models with a single outcome variable:

- $n$ is the number of subjects or individuals in the sample.
- $i$ is the index used for the subjects, thus $i$=1, …, $n$.
- $p$ is the number of predictor variables. The terms independent variable, predictor variable, predictor and feature will be used interchangeably.
- $j$ is the index used for the predictor variables, thus $j$=1, …, $p$.
- $x_{ij}$ is the $j^{\text{th}}$ predictor variable for the $i^{\text{th}}$ subject.
- $x_0$ is a constant for the intercept term, thus $x_{i0} = 1$ for all $i$.
- $\mathbf{X}$ is the $n$ x $(p+1)$ feature matrix containing the intercept and predictor variables for all $n$ subjects

$$\mathbf{X} = \begin{pmatrix} x_{10} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n0} & \dots & x_{np} \end{pmatrix}.$$

- $y_i$ is the observed outcome for the $i^{\text{th}}$ subject.

- $\mathbf{y}$ the outcome vector with length $n$ containing the outcomes $y_i$ for all $n$ subjects

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

- $\hat{y}_i$ is the model prediction for $y_i$ and $\hat{\mathbf{y}}$ is the model prediction for $\mathbf{y}$.
- $\beta_j$ is the model coefficient or model parameter for the $j^{\text{th}}$ predictor variable. $\beta_0$ is the coefficient for the intercept.
- $\boldsymbol{\beta}$ is the coefficient vector with length $(p+1)$ containing the coefficients for the intercept and all $p$ predictor variables

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}.$$

- $\hat{\beta}_j$ is the estimate for $\beta_j$, and $\widehat{\boldsymbol{\beta}}$ is the estimate for $\boldsymbol{\beta}$.

In addition, for longitudinal analyses with repeated outcome measurements on each data point, we use the following notation:

- $m_i$ is the number of measurements per subject, which is variable per individual.
- $N$ is the total number of observations in the dataset,

$$N = \sum_{i=1}^{n} m_i.$$

- $\mathbf{X_i}$ is the $m_i$ x $(p+1)$ feature matrix containing the predictor variables for all $m_i$ observations on the $i^{\text{th}}$ subject

$$\mathbf{X_i} = \begin{pmatrix} x_{10} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{m_i 0} & \cdots & x_{m_i p} \end{pmatrix}.$$

- $\mathbf{y_i}$ is the outcome vector with length $m$ containing the outcomes $y_i$ for all $m_i$ observations on the $i^{\text{th}}$ subject

$$\mathbf{y_i} = \begin{pmatrix} y_1 \\ \vdots \\ y_{m_i} \end{pmatrix}.$$

## 5.2. Statistics versus machine learning

Both statistics and machine learning are tools for data analysis. Although the two fields overlap, there are some distinctions between them. A key difference between traditional statistics and machine learning is in the way data analysis problems are approached. Traditional statistics hypothesises that the observed data were generated by a certain mechanism, and uses the data to test how likely that hypothesis is true. Typically, statistical tests return a *p*-value which is compared to a significance threshold to determine whether or not the data support the hypothesis. The validity of statistical models depends on certain conditions, for example the assumption of normality of the error terms in linear regression. In contrast, machine learning does not make such hypotheses and assumptions but aims to build prediction models by learning from the data what the parameters of those prediction models are (Breiman, 2001a). The usefulness of a machine learning model is assessed by how well the algorithm predicts the outcome in a test dataset. Although it is possible to compute empirical *p*-values, for example through permutation tests, this is not common practice in the machine learning community. Machine learning is often criticised for being a 'black box' solution which does not shed any light on the relationship between predictors and outcome. However, many machine learning algorithms provide variable importance scores or perform variable selection and thus allow distinguishing important predictors from trivial variables.

A practical distinction between the two fields is how they cope with large datasets. Traditional statistical models struggle with large datasets, especially when the number of predictor variables exceeds the number of subjects (the $n < p$ case). As a rule of thumb linear regression requires 10 times as many subjects as there are covariates, whereas logistic regression needs at least 10 events per covariate (Miller & Kunce, 1973; Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996; VanVoorhis & Morgan, 2007). Machine

learning has a strong advantage here as its methods are applicable to high dimensional datasets. For large datasets, such as genetic studies, machine learning allows the simultaneous multivariable analysis of many predictors, which is not possible using traditional statistics.

Statistical models and machine learning algorithms can be broadly grouped into methods for supervised or unsupervised learning. Supervised learning refers to research problems where the aim is to make predictions on a certain outcome variable and a dataset containing the predictor variables and outcome labels is used to train the algorithm. On the other hand, the term unsupervised learning is used for problems where there is no outcome variable and the goal is to detect clusters of similarity in the observations. In this thesis we exclusively use supervised approaches and our discussion on statistical methods is thus limited to supervised models.

## 5.3. Traditional statistical methods

The statistical methods covered in the following paragraphs are linear and logistic regression for cross-sectional data and linear mixed models for repeated measurements.

### 5.3.1. Linear regression

Linear regression is a commonly used method for the analysis of continuous outcome variables. This model relies on the assumption that there is a linear relationship underlying the observed predictor variables and the outcome. Moreover, it is assumed that the residual error terms are random, independent, normally distributed and have constant variance. A linear regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

with the error terms $\varepsilon_i \sim N(0, \sigma^2)$.

When the equations of all $n$ subjects are stacked the model can be written in matrix notation as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon}$ being an $n$-dimensional vector of error terms.

The model parameters are estimated by minimizing the residual sum of squares (RSS), which is the sum of the squared differences between the observed outcomes ($\mathbf{y}$) and the outcome values predicted by the regression model ($\mathbf{X\beta}$).

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)\right)^2 = \|\mathbf{y} - \mathbf{X\beta}\|^2$$

Thus, the least squares estimate for the model parameters is

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}(\|\mathbf{y} - \mathbf{X\beta}\|^2)$$

For linear regression, the same estimates are obtained by maximum likelihood estimation, which returns the parameter values that maximize the likelihood of observing the outcomes $\mathbf{y}$. The significance of parameter estimates, i.e. whether a parameter is different from zero, can be assesed using t-tests and type III F-tests.

### 5.3.2. Logistic regression

Bivariate outcomes can be analysed using logistic regression. Outcome predictions from a linear regression are real numbers, which is not suitable for categorical outcomes. Logistic regression provides a solution to this by modelling the probability of belonging to a certain class. For a binary outcome with levels A and B, logistic regression models the log odds of belonging to class A as opposed to class B, given the observed predictor variables,

$$log\left(\frac{P\left(y_i = A \middle| x_{i1}, x_{i2}, \ldots, x_{ip}\right)}{P\left(y_i = B \middle| x_{i1}, x_{i2}, \ldots, x_{ip}\right)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

As $P(y_i = B|x_{i1}, x_{i2}, \dots, x_{ip}) = 1 - P(y_i = A|x_{i1}, x_{i2}, \dots, x_{ip})$, the probability of belonging to class A is given by

$$P(y_i = A|x_{i1},\ x_{i2}, \dots,\ x_{ip}) = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}}$$

From this equation it is clear that the model appropriately predicts probabilities in the range between 0 and 1.

The predictor variables have a linear effect on the log odds, i.e. for a unit increase in $x_j$ the log odds increases by $\beta_j$. However, the change in the probability $P(y_i = A|x_{i1},\ x_{i2}, \dots,\ x_{ip})$ due to a change in $x_j$ depends on the current value of $x_j$. As there is a linear effect of the predictor variables on a function of the outcome, logistic regression belongs to the class of generalized linear models. No closed form expression for estimating the model parameters $\beta_j$ exists and estimates are obtained by applying iterative algorithms such as the Newton-Raphson method to maximize the likelihood function. Wald tests and likelihood-ratio tests can be used to formally test the statistical significance of predictor variables.

### 5.3.3. Linear mixed models for longitudinal repeated measures

Some studies, for example clinical trials, follow a group of subjects over a period of time and make repeated observations. A linear regression will not suffice to analyse longitudinal data as it does not take the sequential nature of the observations into account nor the fact that observations on the same subject are likely to be correlated. Linear mixed models capture the correlation between multiple outcome observations and can accommodate unbalanced data, both in terms of the number of observations per subject ($m_i$) as well as the measurement time points. For the context of clinical trials, this is a useful property as patients who dropped out of the study or missed appointments can still be included in the analysis. Assuming that any missing data points are missing at random, the model provides

valid inference on the model parameters (Molenberghs et al., 2004; Verbeke & Molenberghs, 2000).

Linear mixed models are multivariate regression models that model population-level effects and at the same time allow subject-specific deviations from the average trend. The regression terms that capture the average trend in the population are the fixed effects, whereas the subject-specific deviations are modelled by the random effects. Typically, random intercepts and random time effects are considered when modelling longitudinal data.

For $n$ subjects, where the number of data points per subject ($m_i$) is variable, a linear mixed model with $p+1$ fixed effects (the intercept and $p$ predictor variables) and $q$ random effects is specified by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

where $\mathbf{Y}_i$ is a $m_i$ dimensional vector of longitudinal outcome observations on the $i^{\text{th}}$ subject, $\mathbf{X}_i$ is a $m_i$ x $(p+1)$ dimensional matrix of fixed predictor variables, $\mathbf{Z}_i$ is a $m_i$ x $q$ dimensional matrix of random predictor variables, $\boldsymbol{\beta}$ is a $(p+1)$ dimensional vector of fixed effects parameters, $\mathbf{b}_i$ is a $q$ dimensional vector of subject specific random effects parameters and $\boldsymbol{\varepsilon}_i$ is a vector of residual error terms. Furthermore, $\mathbf{b}_i \sim N(0, \mathbf{D})$ where $\mathbf{D}$ is a $q$ x $q$ dimensional variance-covariance matrix and $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\Sigma}_i$ is a $m_i$ x $m_i$ dimensional variance-covariance matrix.

The fixed effects model the average trend in the data, whereas the random effects account for the correlation between observations within the same subject through the variance - covariance matrix $\mathbf{D}$, and any remaining variability is captured in the residual variance-covariance matrix $\boldsymbol{\Sigma}_i$. The model parameters can be estimated using maximum likelihood (ML) or restricted maximum likelihood (REML) (Verbeke & Molenberghs, 2000).

Often interest lies mainly in the estimation of the fixed effects parameters and the random effects are regarded as nuisance parameters. In that case, the marginal perspective of the linear mixed model is sufficient for inference on the fixed effect model parameters. The outcomes are assumed to follow a normal distribution with the fixed effects defining the mean, and where the random effects only appear in the variance term,

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta},\ \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i)$$

Two types of standard errors can be calculated for the fixed effects parameter estimates. Model-based standard errors depend on the variance of $\mathbf{Y}_i$, $Var(\mathbf{Y}_i)$, and assume that the covariance matrix of the marginal model is correctly specified. However, choosing the right random effects structure, i.e. which random effects to include in the model and their covariance matrix, is not straightforward. Mistakes in the random effects structure invalidate inference relying on model-based standard errors. Alternatively, empirical standard errors can be used which replace $Var(\mathbf{Y}_i)$ by the squared residuals $\left(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}\right)^2$. The advantage of empirical standard errors is that they are robust against misspecification of the random effects structure given that the fixed effects are correctly defined (Liang & Zeger, 1986; Verbeke & Molenberghs, 2000). Including the right random effects and specifying their covariance matrix correctly is thus not essential for valid inference, although it improves efficiency of the model. As in the case of linear regression, fixed effects parameters can be formally tested for significance using t-tests and type III F-tests.

## 5.4. Machine learning

In the next paragraphs, some key concepts of machine learning are introduced. We describe the use of independent training and test data and different measures of predictive performance. Next, cross-validation is presented as a method to optimize the hyperparameters of machine learning algorithms and the practice of feature selection to

reduce the dimensionality of the data is explained. Finally, the machine learning methods applied in this thesis, regularised regression, tree-based algorithms and support vector machines, are introduced.

### 5.4.1. Independent training and test data

Predictive machine learning algorithms are evaluated on their predictive performance. To obtain an unbiased estimate of predictive ability, an algorithm is tested on a dataset independent of the data used to build the model. It should be noted however that all variables present in the training data are also required in the test data. In the absence of an independently collected test dataset, the available data can be randomly split in two subsets, one of which is used for algorithm training and the other set aside for measuring test set predictive performance. In this case, the number of observations used for training the model and the number used for accurate prediction assessment are balanced against each other. Commonly used proportions for data splitting are 70 - 80% for training and 20 - 30% for testing. However, this is an arbitrary choice and the available sample size should be considered.

### 5.4.2. Measuring predictive performance

A wide range of statistics can be used to measure predictive ability and compare different machine learning algorithms. Optimizing the predictive performance of an algorithm is equivalent to minimizing the prediction error. The measure used depends on the type of outcome variable and should be decided upon before starting model optimization and selection. In the next paragraphs the measures used in this thesis are introduced, but this list is by no means exhaustive. Depending on the research question, ad hoc performance measures may be defined.

*5.4.2.1. Continuous outcome*

When predicting a continuous outcome, a widely used statistic to assess predictive performance is the root mean squared error (RMSE). The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

and lower RMSE scores reflect better prediction. The mean squared error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is an equivalent statistic.

Another popular measure for predicting continuous outcomes is the $R^2$ statistic. The $R^2$ is given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of $y_i$ across the sample

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i \,.$$

This statistic expresses the proportion of variability in the outcomes that is explained by the model. By definition the $R^2$ ranges between 0 and 1 and higher values correspond to better prediction.

*5.4.2.2. Binary outcome*

For binary outcome variables, a 2x2 confusion matrix contrasts observed outcomes with predictions (Table 3). The accuracy is the proportion of observations that are correctly classified,

$$\text{accuracy} = \frac{\text{True positives} + \text{True negatives}}{n}.$$

The accuracy can also be extended to the context of a multi-class categorical outcome.

Other measures that can be derived from the confusion matrix are sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). However, these statistics are not suitable for assessing the predictive ability of an algorithm. While sensitivity and specificity can reach 100% by classifying every data point as positive or negative, respectively, the PPV and NPV focus on one outcome class only. These classification measures are examined in more detail in part 1 of the thesis in the context of a PGx test for clozapine induced agranulocytosis.

When the prediction algorithm outputs a probability rather than a classification decision, observations can be classified as positive when the predicted probability exceeds a specified threshold and negative otherwise. A receiver-operator characteristic (ROC) curve plots 1-specificity against sensitivity for varying decision threshold values (Fig. 9). The area-under-the-curve (AUC) is a measure of how well the model predicts and ranges between 0 and 1. High AUC scores indicate better prediction, while a value of 0.5 corresponds to a random chance classifier.

Table 3. Confusion matrix for the prediction of a binary outcome with levels *positive* and *negative*.

| **Observed outcome** | **Predicted outcome** | |
|---|---|---|
| | Positive | Negative |
| Positive | True positives | False negatives |
| Negative | False positives | True negatives |

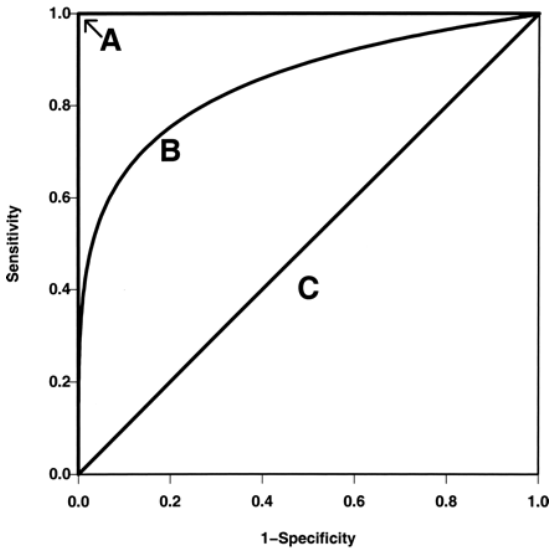



Figure 9. ROC curves of three hypothetical prediction models. Model A is a perfect classifier with AUC=1, model B has an AUC=0.85 and model C classifies randomly resulting in AUC=0.5. Adapted from Zou et al. (2007).

### 5.4.3. Validation and cross-validation

Many machine learning algorithms depend on one or more tuning parameters or hyperparameters. Choosing the tuning parameter value that leads to optimal prediction in the training data likely overfits the algorithm to the training data. To increase generalizability, an independent validation dataset can be used select to the hyperparameter value that maximizes prediction accuracy (Fig. 10). Given a list of possible values for the hyperparameter, the training data are used to construct models for each value in that list. The algorithm that achieves the best prediction in the validation data is selected and in a final step an unbiased estimate of predictive performance is obtained by applying the algorithm to a test dataset. As the training data, validation data and test data are independent of each other this strategy requires a large total sample size.

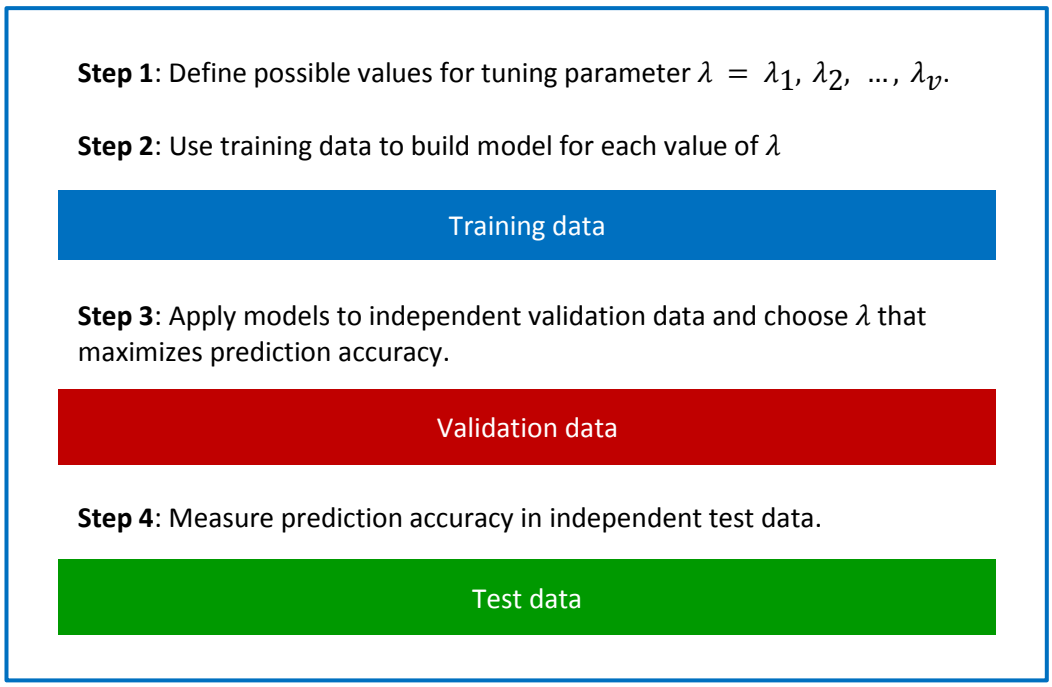



Figure 10. Hyperparameter tuning using validation data.

Alternatively, cross-validation on the training data can be used to set the value of tuning parameters without need for a validation dataset. Again, a list of possible values for the hyperparameter to be tuned is defined. Then, for $k$-fold cross-validation the training data is split in $k$ equally sized subsamples or folds. Each fold is used to test how well a model fitted on the remaining $k$-1 folds performs (Fig. 11). This step is repeated for each possible value of the tuning parameter and the value that leads to the best prediction performance averaged across all folds is selected. Finally, the algorithm is fit on the entire training data with the hyperparameter set to the selected value and the prediction of this model is measured in independent test data. When the machine learning algorithm has multiple hyperparameters, a grid of tuning values is cross-validated to select the optimal combination.

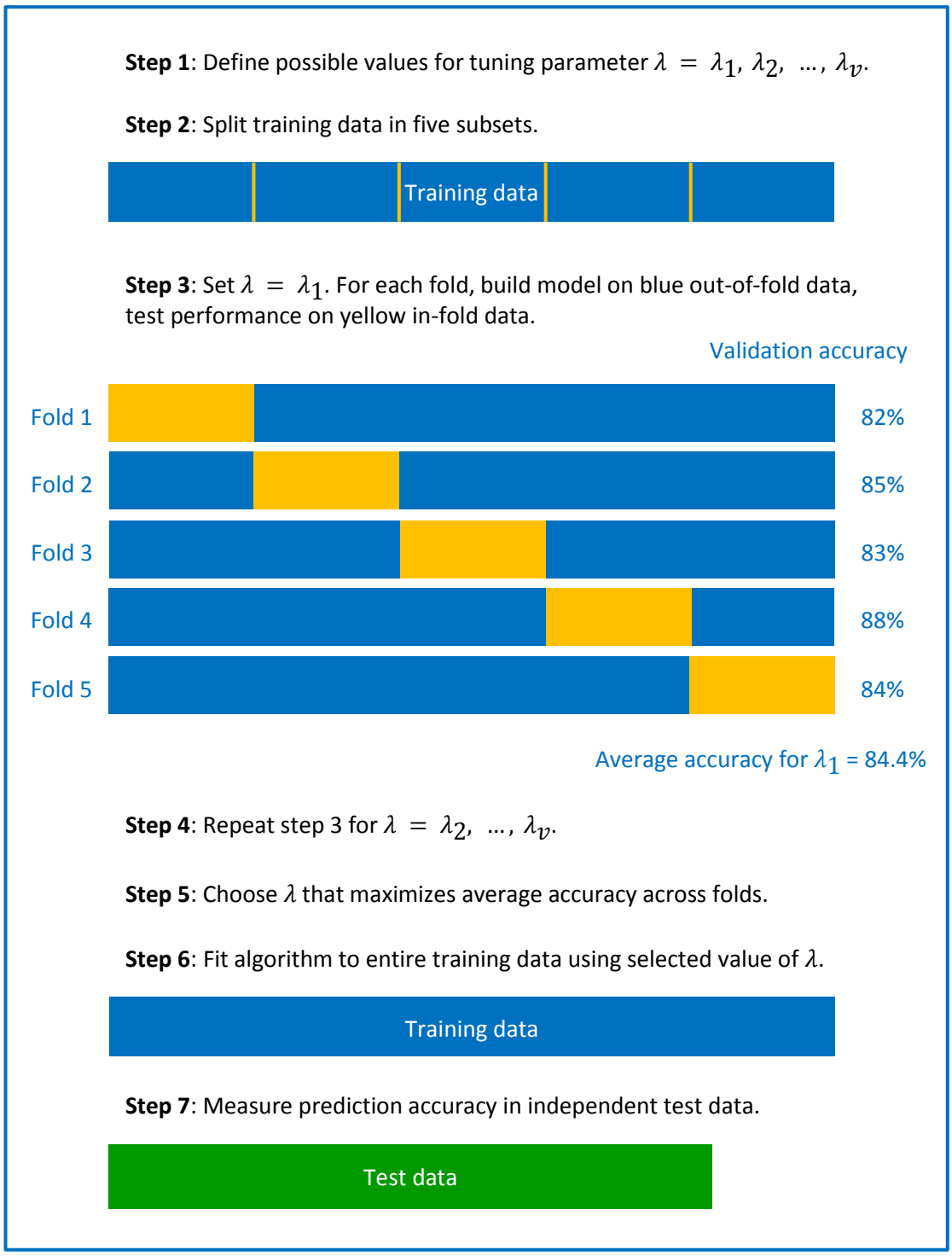



Figure 11. Hyperparameter tuning using five-fold cross-validation.

As cross-validation requires repeatedly fitting the model to a subset of the data, it is computationally expensive. Typical values for $k$ are 5 and 10, though a different number of folds may be used balancing the computational cost of repeating the analysis $k$ times and

the size of each fold. Moreover, smaller values of $k$ result in less bias but more variance in the predictive power estimate (James, Witten, Hastie, & Tibshirani, 2013). The $k$-fold cross-validation process can be repeated a certain number of times, which is more robust but evidently even more computationally expensive. A special case of $k$-fold cross-validation is leave-one-out cross-validation, where $k$ equals the number of samples in the training data.

For linear models like lasso and elastic net, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used for hyperparameter tuning as an alternative to cross-validation (Ninomiya & Kawano, 2016). These statistics are calculated on the entire training sample and do not require repetitive model fitting. Asymptotically, AIC is equivalent to leave-one-out cross-validation and BIC to leave-$m$-out cross-validation, where $m$ is a function of the sample size (Shao, 1997; Stone, 1977).

#### 5.4.4. Feature selection

In large datasets with many predictor variables, it is likely that only a subset of features contribute to prediction whereas other features are irrelevant. Variable selection or feature reduction removes these redundant predictors and thus leads to more sparse algorithms. Moreover, removing noise variables from the model can improve prediction accuracy and avoid overfitting (Guyon & Elisseeff, 2003; Saeys, Inza, & Larrañaga, 2007). In addition, identifying the subset of relevant predictors may increase understanding of the biological mechanisms of the phenotype studied. Three main classes of feature reduction methods exist (Saeys et al., 2007). The simplest technique is applying a filter method. The features are ranked using parametric or non-parametric tests and the top scoring variables are carried forward as features in the prediction model. Machine learning methods such as random forest and SVM can even be used to rank and select features prior to fitting a separate machine learning model. Filter methods are fast and scalable but independent of the prediction algorithm. Secondly, wrapper techniques can be applied. Here, the

prediction algorithm is incorporated in the variable selection process. For example, recursive feature elimination (RFE) iteratively removes features by repeatedly fitting an algorithm excluding one variable at a time. Then, the variable with the least impact on model performance is excluded from the feature set. This process is repeated until all variables are exhausted. Wrapper methods take dependencies between the features into account but quickly become computationally expensive as the number of features grows. Lastly, some machine learning methods have embedded feature selection properties. For example elastic net and decision trees inherently perform variable selection. Embedded variable selection is computationally efficient but not available for all machine learning algorithms. As for all model building steps, feature reduction should be carried out on the training data only.

### 5.4.5. Regularized regression models

Traditional linear regression models run into overfitting problems when modelling datasets with many correlated variables, but penalized regression - a machine learning extension of generalized linear regression - provides a way around this problem.

The expected error of linear regression predictions, i.e. the squared difference between the estimated outcome and the true outcome, can be decomposed into

$$\mathrm{E}[(y_i - \hat{y}_i)^2] = \sigma^2 + Bias^2 + Variance = \sigma^2 + MSE.$$

The variance in prediction estimates is due to the randomness in the training data, whereas the bias reflects the deviation of the average predictions from the true average (Fig. 12) (Hastie, Tibshirani, & Friedman, 2009). $\sigma^2$ is the irreducible error of the test set outcomes. The MSE of the predictions is the sum of the squared bias and the variance.

Maximum likelihood estimates for $\beta_j$ are unbiased and the variance is the only factor contributing to the MSE. However, sometimes accepting a small bias can lower the MSE by reducing the variance. This balance is referred to as the bias-variance trade-off.

As in the case of linear regression, a least squares procedure is used to optimize model coefficients ($\boldsymbol{\beta}$), but penalized regression models introduce some bias by adding a penalty term to the RSS.

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + penalty)$$

Here, the concept of penalized regression is explained in a linear regression context but it can easily be extended to logistic regression and other generalized linear models by maximizing the penalized log-likelihood (Hastie et al., 2009).

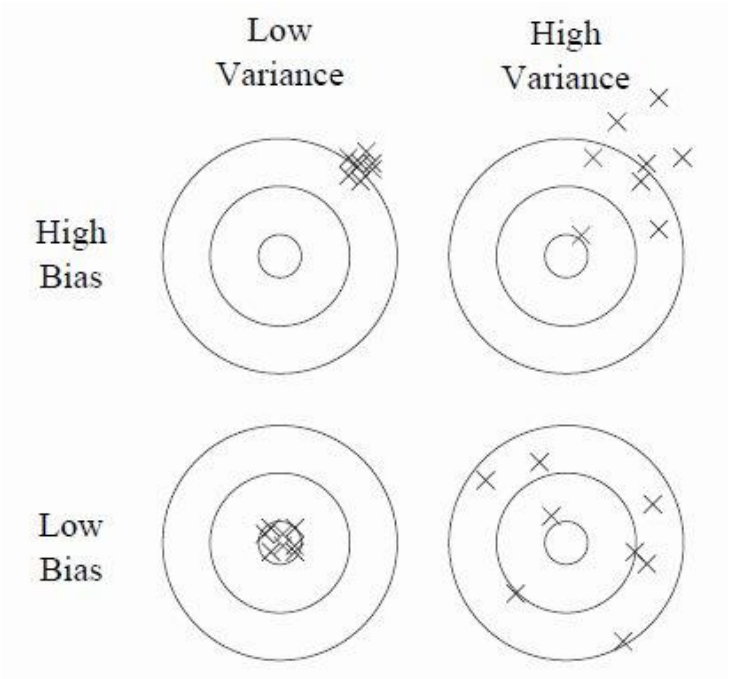



Figure 12. Bias and variance of predictions. Crosses represent the predicted values, whereas the true mean is located in the centre of the circles. Adapted from Domingos (2012).

*5.4.5.1. Lasso*

In the lasso model (*l*east *a*bsolute *s*hrinkage and *s*election *o*perator), the residual sum of squares is minimized subject to the $l_1$-norm or lasso penalty on the absolute size of the coefficients. The lasso model coefficient estimates are obtained by

$$\widehat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right).$$

In the above equation, $\lambda$ is a tuning parameter controlling the strength of the penalty. The value of $\lambda$ is positive and can be set by cross-validation.

The lasso optimization problem can also be rewritten as

$$\widehat{\boldsymbol{\beta}}^L = \arg\min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2) \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

where $s$ has a one-to-one relationship with $\lambda$. Larger values of $\lambda$ correspond to smaller values of $s$.

The lasso shrinks the coefficients towards zero and sets some of them equal to exactly zero, thereby in effect dropping them from the model. The lasso thus automatically performs feature selection. The larger the value of $\lambda$ is, the more sparse the lasso model is. A disadvantage of lasso is that the model only retains one predictor from a set of correlated variables while excluding the rest. In addition, in the $n < p$ case, the lasso is restrained by the sample size of the training data and can select a maximum of $n$ features (H. Zou & Hastie, 2005). The left panel of Figure 13 shows a graphical representation of the lasso coefficient optimization for a model with two features. The least squares or ML estimate $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ is marked on the plot. The elliptical contours around $\widehat{\boldsymbol{\beta}}$ represent values of $\beta_1$ and $\beta_2$ that result in the same RSS, while the blue area is the region where $\beta_1$ and $\beta_2$ comply with the lasso constraint $|\beta_1| + |\beta_2| \leq s$. Thus, the lasso estimate $\widehat{\boldsymbol{\beta}}^L$ is the point in

the blue diamond with the lowest RSS. In this example, the lasso model selects only the first variable and excludes the second by setting $\beta_2$ equal to 0.

*5.4.5.2. Ridge regression*

Ridge regression applies an $l_2$-norm or ridge penalty to the coefficient least squares procedure, restraining the size of the coefficients. The ridge estimates are obtained by

$$\widehat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \right).$$

Again, $\lambda$ is a tuning parameter to be optimized by cross-validation. Larger values of $\lambda$ impose a stronger penalty, thus coefficients are more shrunk towards zero.

In analogy with the lasso penalty, the ridge regression optimization can be expressed as

$$\widehat{\boldsymbol{\beta}}^R = \arg\min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2) \text{ subject to } \sum_{j=1}^{p} {\beta_j}^2 \leq s.$$

Ridge regression shrinks the coefficients towards zero and each other. Unlike lasso, ridge regression does not set any coefficient to zero and thus does not perform variable selection. All features are retained in the model. The estimated coefficients of correlated variables have similar sizes, as if the effect is spread across the correlated predictors. The right part of Figure 13 symbolizes ridge regression coefficient optimization for a model with two features. The blue circle indicates the area where ${\beta_1}^2 + {\beta_2}^2 \leq s$. The ridge estimate $\widehat{\boldsymbol{\beta}}^R$ is the point within the blue circle with the lowest RSS.

For both lasso and ridge regression, the size of the blue constraint area depends on the value of $\lambda$. Smaller values of $\lambda$ result in larger blue regions, thus less penalization. When $\lambda$ is sufficiently small, $\widehat{\boldsymbol{\beta}}$ is included in the constraint area and $\widehat{\boldsymbol{\beta}}^L$ and $\widehat{\boldsymbol{\beta}}^R$ will be equal to the least squares estimate. In the case of $\lambda = 0$, ridge regression and lasso reduce to a traditional linear regression model.

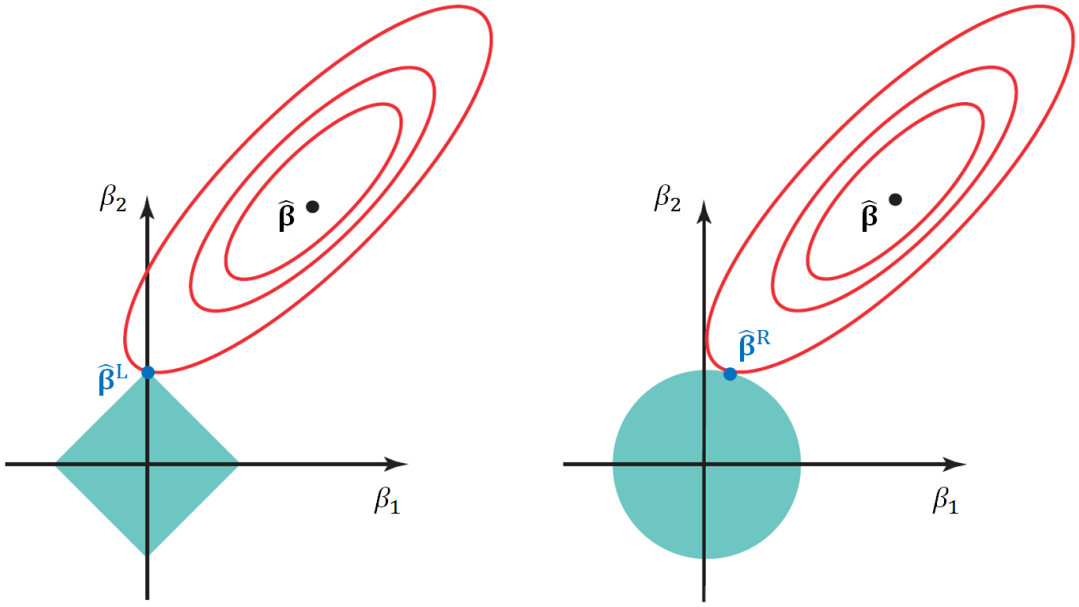



Figure 13. Coefficient estimation subject to lasso (left) and ridge (right) penalty. $\widehat{\boldsymbol{\beta}}$ indicates the location of the least squares estimate. The red ellipses represent values of $\boldsymbol{\beta}$ that result in constant RSS. The blue shapes are the regions were $\boldsymbol{\beta}$ complies with the constraints of $|\beta_1| + |\beta_2| \leq s$ for lasso and $\beta_1{}^2 + \beta_2{}^2 \leq s$ for ridge regression. $\widehat{\boldsymbol{\beta}}^L$ indicates the lasso estimate and $\widehat{\boldsymbol{\beta}}^R$ the ridge estimate. Adapted from James et al. (2013).

*5.4.5.3. Elastic net*

The elastic net is a penalized regression model that combines the advantages of lasso and ridge regression. The elastic net performs grouped feature selection, retaining important variables while excluding irrelevant predictors, and selects correlated variables in or out of the model as a group (Zou & Hastie, 2005). Unlike lasso, the number of features that can be selected is not limited by the size of the training sample.

The elastic net penalty is a weighted average of the $l_1$ and $l_2$-norm. The model coefficients are estimated by

$$\widehat{\boldsymbol{\beta}}^{EN} = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \left( \alpha|\beta_j| + (1-\alpha)\beta_j{}^2 \right) \right).$$

The elastic net has two hyperparameters: $\lambda$ determines the strength of the penalty whilst $\alpha$ is limited to values between 0 and 1 and is a mixing factor between the $l_1$ and $l_2$-norm. Cross-validation can be used to select the values for $\alpha$ and $\lambda$.

It is clear that ridge regression and lasso are special cases of elastic net when $\alpha = 0$ and $\alpha = 1$, respectively. A visual comparison of the elastic net constraint area with the lasso and ridge regression is shown in Figure 14. The curvature of the elastic net constraint area depends on $\alpha$, whereas the size is determined by $\lambda$.

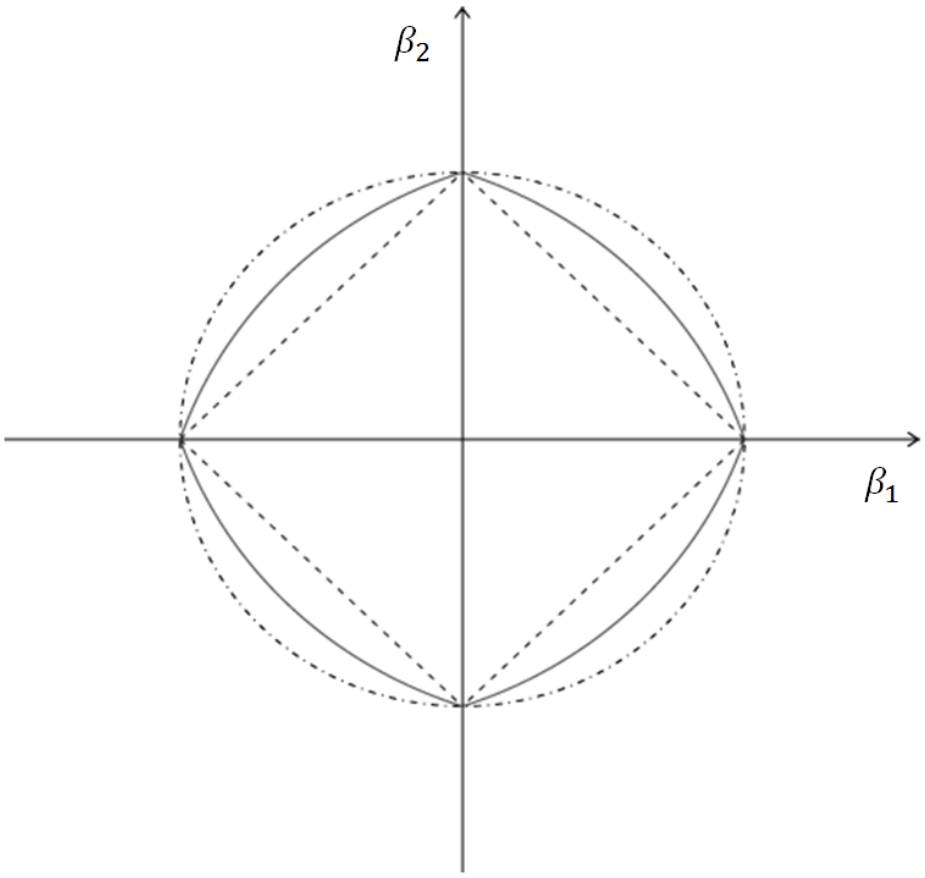



Figure 14. Contours of the constraint area for model coefficient estimates. - · - · - · - indicates the shape of the ridge penalty, - - - - - - of the lasso penalty and ——— of the elastic net penalty with $\alpha = 0.5$. Adapted from Zou and Hastie (2005).

*5.4.5.4. Penalized linear mixed models*

As the $l_1$ and $l_2$ penalties can be added to the log-likelihood of a model for penalized coefficient estimation, it follows that penalized regression can also be extended to the multivariate case. Such models are of interest for the analysis of longitudinal data with a large number of predictor variables. We discuss the concept of penalized linear mixed models without going into technical details. Although longitudinal statistical learning is still an emerging field, a few penalized linear mixed models have been developed (Table 4). These models vary in the penalty applied to the coefficient estimation and the estimation algorithm that is used. Furthermore, feature selection can be performed on the fixed or random effects, or on both at the same time (Chen, Grant, Wu, & Bowman, 2014). The penalized linear mixed models listed here are all slightly different models and each method can be fitted with the appropriate software.

For the longitudinal analysis of an antidepressant clinical trial (chapter 8) we used the lassop software to estimate a linear mixed elastic net, because it permits an elastic net penalty to be imposed on a linear mixed model (Rohart, 2016; Rohart, San Cristobal, & Laurent, 2014). When fitting this model, the random effects are treated as missing values and the ML estimation of the fixed effects is subject to an elastic net penalty. A limitation of the lassop software is that the model can select at maximum the same number of predictors as there are observations in the data. The software can thus not fit a ridge regression, which retains all predictor variables, when the number of predictors exceeds the number of subjects (the $n < p$ case). In $n < p$ situations REML cannot be used to estimate model coefficients and thus ML is the only estimation procedure supported by the lassop software (Agbedjro, 2017).

Table 4. Software packages for penalized linear mixed model analysis.

| **Method** | **Penalty** | **Variable selection on** | **Suitable for *n* < *p*** | **Software** | **Reference** |
|---|---|---|---|---|---|
| lassop | Elastic net | Fixed effects | Yes | R package: MMS | (Rohart et al., 2014) |
| lmmlasso | Lasso | Fixed effects | Yes | R package: lmmlasso | (Schelldorfer, Bühlmann, & Van De Geer, 2011) |
| glmmixedlasso | Lasso | Fixed effects | Yes | R package: glmmixedlasso | (Schelldorfer, Meier, & Bühlmann, 2014) |
| glmmLasso | Lasso | Fixed effects | No | R package: glmmLasso | (Groll & Tutz, 2014) |
| Pen.LME | Adaptive lasso | Fixed and random effects simultaneously | No | R code | (Bondell, Krishna, & Ghosh, 2009; Bondell, Krishna, & Ghosh, 2010) |
| Indep. selection | Elastic net | First fixed effects, then random effects | No | MATLAB | (Fan & Li, 2012) |

### 5.4.6. Tree-based prediction algorithms

*5.4.6.1. Decision trees*

When the relationship between predictor variables and the outcome is non-linear, regression models fail to fit this adequately, and tree-based machine learning methods perform better (Fig. 15).

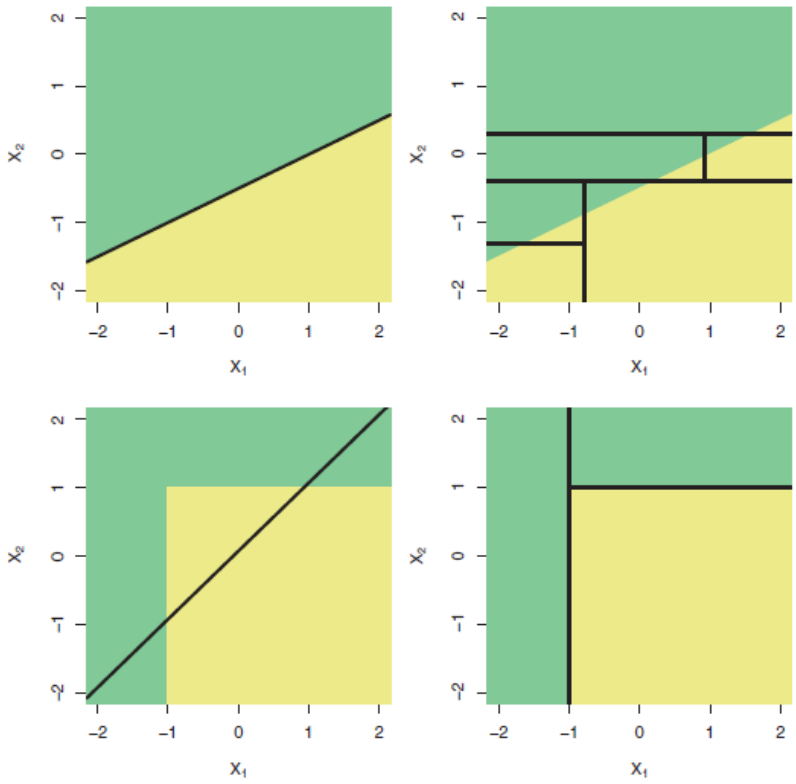



Figure 15. Top row: A binary classification example where the true decision boundary is linear. A linear model (left) captures the relationship perfectly but a decision tree (right) struggles to classify the two categories. Bottom row: When the true decision boundary is non-linear a decision tree (right) can distinguish the classes whereas the linear model (left) fails to do so. Adapted from James et al. (2013).

A decision tree is an algorithm that recursively splits observations using the values of predictor variables (Fig. 16). Decision trees are greedy algorithms, because at each step the best split is considered without taking future splitting steps into account. Split points in the

tree are referred to as internal nodes, whereas the subsets at the end of the branches are external nodes or leaves. At each node, the feature space is split in two subregions based on the value of a predictor variable. Continuous features can separate the predictor space by splitting on a value $t$. All observations with $x_j \leq t$ are assigned to the left subregion whereas all observations with $x_j > t$ go to the right subregion. Categorical splitting variables place observations with certain classes in one subregion and the remaining classes in the other subregion. In Figure 16, predictors $x_1$and $x_2$ are continuous, whereas $x_3$is a categorical predictor with three classes. Predictor variables can be used in more than one node of the decision tree, for example two distinct values of a continuous predictor can be used as split points in different nodes.

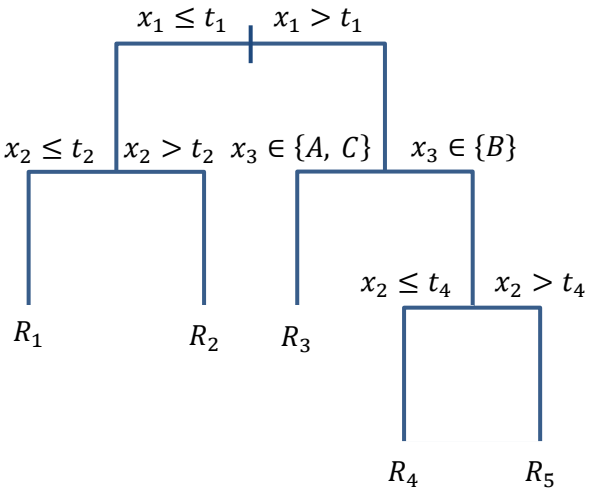



Figure 16. Example of a decision tree with four internal nodes where the sample is split, resulting in five terminal nodes corresponding to five subregions in the predictor space ($R_1$ to $R_5$). $x_1$ and $x_2$ are continuous predictors, $x_3$ is a categorical predictor with classes A, B and C.

A decision tree can be applied to predict categorical (classification tree) as well as continuous outcomes (regression tree) and in both cases the algorithm optimizes the purity of the nodes. Classification trees aim to separate the outcome categories by searching for

the predictor variable that leads to the cleanest split in each node. Purity of a node can be measured by the classification error rate

$$E = 1 - \max_k \hat{p}_k.$$

where $\hat{p}_k$is the proportion of training observations that belong to the $k^{\text{th}}$ class, for a categorical outcome with $K$ classes. Two alternative measures that are more sensitive than the classification error are the Gini index

$$G = \sum_{k=1}^{K} \hat{p}_k(1 - \hat{p}_k)$$

and the cross-entropy

$$C = -\sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k.$$

Smaller values of $G$ and $C$ indicate a more pure split in the node and thus the algorithm aims to minimize these quantities.Classification trees make predictions using the most common class seen in the training data in each terminal node. Predictions for a test sample are made by following the branches of the classification tree and placing the new observation is a terminal node. The predicted class for the test observation is the most common class in the training data in that terminal node.

Regression trees aim to minimize the RSS between the observed outcomes and the node predictions. For a continuous outcome $y$, a piecewise constant regression tree splits the training data in subregions and models the outcome by taking the mean of $y$ within each node. The predictor values used for splitting the nodes are chosen so that RSS in the subregions is minimized. Some decision tree algorithms can fit more complex regression models in the nodes. For example, the Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) decision tree algorithm fits a regression model in each node and the residuals are grouped by positive or negative value. Then, the predictor variable that results

in the best split between the two groups of residuals is selected. The regression models within the nodes can be constant models (intercept only), simple or multiple linear regression models. To make predictions using a regression tree, a new observation is run through the tree until it reaches a terminal node. The predicted outcome is the mean outcome of the training data in that terminal node.

In both classification and regression decision trees, splitting of the nodes is repeated until the terminal nodes contain less than a pre-specified number of observations. To prevent overfitting the tree to the training data, decision trees can be pruned, i.e. recursively removing the least important node from the tree. The appropriate number of nodes is a tuning parameter that can be determined using cross-validation.

An advantage of decision trees is that they are easily interpretable and can be graphically displayed. Moreover, decision trees perform variable selection, except in the scenario where all predictor variables are used in the splitting process. Although the simplicity of decision trees makes them attractive prediction models, their predictive performance is generally lower than that of more sophisticated algorithms such as random forest (James et al., 2013).

*5.4.6.2. Random forest*

The random forest algorithm aggregates multiple decision trees, which can substantially improve the predictive power compared to a single decision tree (Breiman, 2001b). A bootstrapped sample of the training data is drawn and used in the construction of each individual tree (Fig. 17). Furthermore, for each tree only a random subset of predictors is considered. Typically, the number of predictor variables considered in the construction of each tree ($r$) is $\sqrt{p}$ for classification forests and $p/3$ for regression forests, though $r$ is a hyperparameter that can be tuned (Hastie et al., 2009). The randomness introduced when building the individual trees assures that the trees in the random forest are uncorrelated.

For prediction, a new observation is run through each of the decision trees in the random forest and the predicted outcome is obtained by majority-voting (classification) or averaging across trees (regression). Averaging across trees reduces the prediction variance, which is further reduced by the low correlation between the trees (Hastie et al., 2009). A useful feature of random forests is the fact that bootstrapping the training data creates out-of-bag (OOB) samples. These can be used as validation data to get a prediction estimate: for each observation a prediction is made using only trees for which that observation was not included in the bootstrapped sample. The OOB prediction estimate provides an alternative to cross-validation without any additional computational cost. Furthermore, the gain in OOB predictive performance can be monitored as trees are added to the random forest. Once the OOB prediction accuracy or the OOB prediction error stabilizes the random forest training can be stopped. An additional advantage is that adding more trees to the forest does not overfit the model, thus the size of the random forest is a hyperparameter that does not need to be tuned. The degree of pruning applied to the individual trees can be optimized, though in practice this parameter does not have a large impact on the model performance (Hastie et al., 2009). Importance scores for predictor variables can be calculated by averaging the reduction in RSS or Gini index across all trees in the random forest. Alternatively, the OOB samples can be used to construct importance scores by permuting the values of a single predictor variable and measuring the reduction in prediction accuracy. A special case of random forest is bagging, where $r = p$. In bagging, all predictors are considered in the construction of each decision tree and by consequence the trees are correlated. For datasets where the proportion of relevant predictors is small, bagging can outperform random forest. In that case the probability of including a relevant predictor in the fraction $r$ is small and thus many trees in a random forest consist of noise variables only.

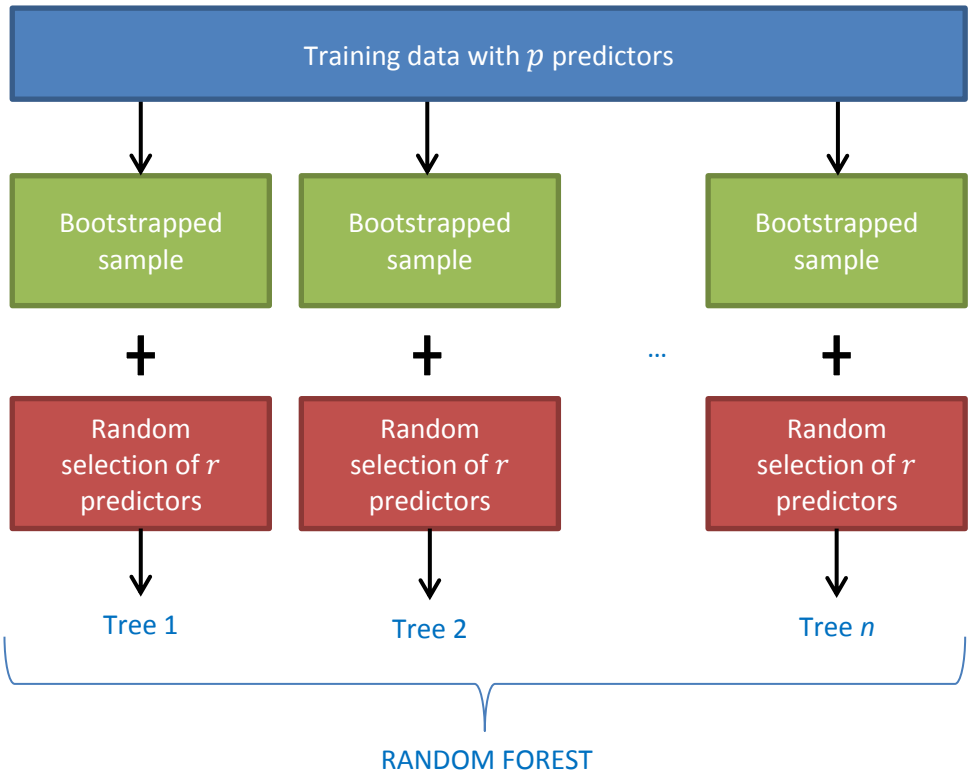



Figure 17. Graphical representation of the random forest algorithm.

### 5.4.7. Support vector machine

The support vector machine (SVM) is a powerful machine learning technique for binary classification. The SVM searches for the hyperplane in high dimensional feature space that separates the outcome classes with the largest margin. A linear SVM uses the data in $p$-dimensional space, where each dimension represents one of $p$ predictor variables. Each observation is thus a vector in the feature space. The observations situated closest to the separating hyperplane are called the support vectors (Fig. 18). The decision boundary is found by maximizing the margin, i.e. the perpendicular distance from the support vectors to the hyperplane. This means that the location of the decision boundary depends only on the support vectors and not the remaining data points.

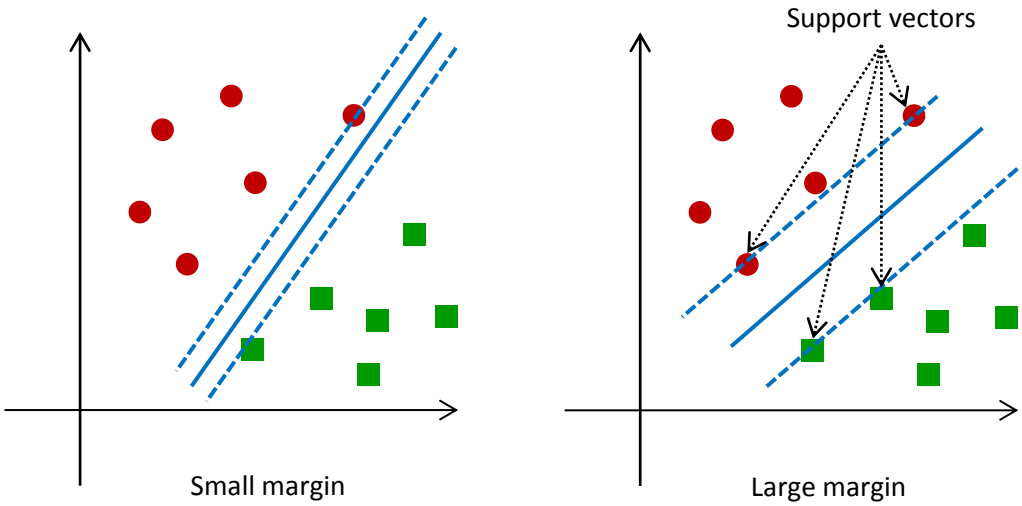



Figure 18. In the left panel, the red and green classes are separated by a small margin. In the right panel, the margin is large.

*5.4.7.1. Perfect linear separation*

We first introduce the SVM for the scenario of a perfectly linearly separable classification problem (Fig. 19). For a binary outcome with two classes $y_i \in \{-1, 1\}$ a hyperplane is defined as

$$H(x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = 0.$$

The observations are classified by their location relative to the hyperplane, thus

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0 \text{ if } y_i = 1$$

and

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0 \text{ if } y_i = -1.$$

This is equivalent to

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0.$$

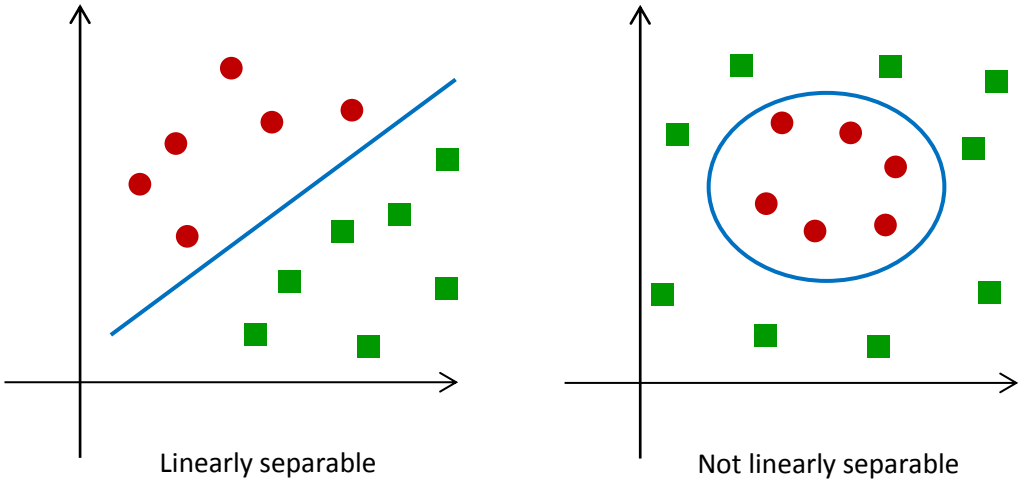



Figure 19. In the left panel, the red and green classes are linearly separable. The right panel shows an example of non-linearly separable data.

The distance between a data point $x_i$ (a $p$-dimensional vector) and a hyperplane $H(x)$ is given by

$$\frac{H(x_i)}{\|\boldsymbol{\beta}\|}.$$

with $\|\boldsymbol{\beta}\| = \sum_{j=1}^{p} \beta_j^2$. Note that the intercept $\beta_0$ is excluded in $\|\boldsymbol{\beta}\|$.

The optimal separating hyperplane maximizes the margin $M$ between the hyperplane and the support vectors

$$\frac{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{\|\boldsymbol{\beta}\|} \geq M$$

where $M > 0$. However, given the hyperplane $H(x) = 0$, every $kH(x) = 0$ with $k \neq 0$ is also a hyperplane. Therefore, to uniquely define the separating hyperplane we add the constraint that

$$\|\boldsymbol{\beta}\| = \frac{1}{M}.$$

Thus, maximizing $M$ is equivalent to minimizing $\|\boldsymbol{\beta}\|$ subject to $y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq 1$, which is a convex optimization problem (Hastie et al., 2009). It can be derived

using Lagrange multipliers that the solution for $\boldsymbol{\beta}$ is a linear combination of the support vectors. The location of the hyperplane thus only depends on the data points closest to it.

It is useful to note that the only term in the Lagrange function that uses the training data consists of the inner product between the observations. The inner product between two data points $x_a$ and $x_b$ is defined as

$$\langle x_a, x_b \rangle = \sum_{j=1}^{p} x_{aj} x_{bj}.$$

This feature will prove helpful when we discuss non-linear SVMs.

*5.4.7.2. Non-separable classification*

In situations where it is not possible to perfectly separate outcome classes by a hyperplane, it is useful to allow some flexibility. To tolerate some observations to fall in the margin or even on the wrong side of the hyperplane, the maximum margin equation can be modified to

$$\frac{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{\|\boldsymbol{\beta}\|} \geq M(1 - \epsilon_i)$$

where $\epsilon_i = 0$ if the $i^{\text{th}}$ observation is on the correct side of the hyperplane and outside of the margin, $1 > \epsilon_i > 0$ if it falls on the correct side but inside the margin and $\epsilon_i > 1$ if it is on the wrong side of the hyperplane. Furthermore,

$$\sum_{i=1}^{n} \epsilon_i \leq C$$

where *C* is the cost-parameter controlling how strictly the separation between the classes is enforced. A hard margin ($C = 0$) requires all items to be correctly classified, whereas a soft margin (large $C$) allows some objects to be misclassified (Cortes & Vapnik, 1995). Allowing some flexibility increases the generalizability of the algorithm to test data.

*5.4.7.3. Non-linear separation*

Not all classification problems can be solved using a linear separating hyperplane. As an alternative to linear separation, the input data can be projected to higher dimensional space (Fig 20). In fact, because the optimization function only depends on the inner product of the training data, the transformation function does not need to be known explicitly. A kernel function computing the inner product of the transformed data points is sufficient,

$$K(x_a, x_b) = \langle h(x_a), h(x_b) \rangle$$

where $h(x)$ is the transformation function applied to the input data. This is known as the *kernel trick*. Popular kernel functions are radial and polynomial kernels

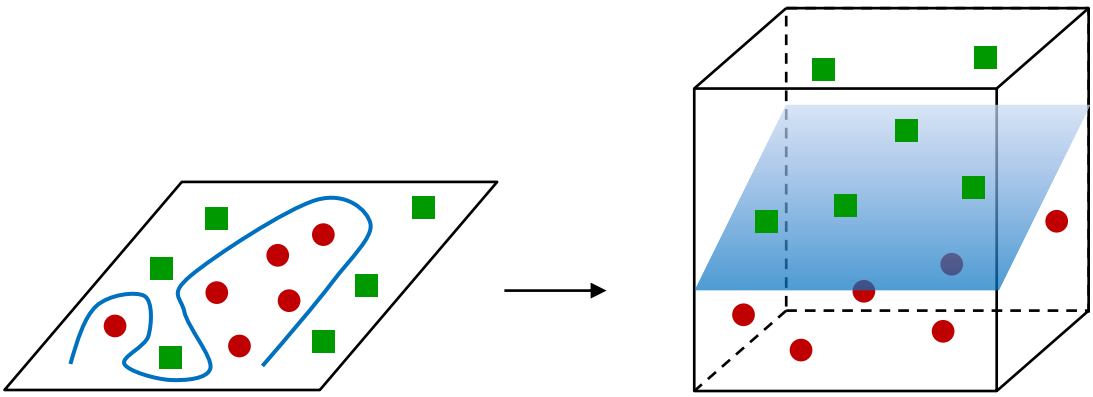
$$Radial\colon K(x_a, x_b) = \exp(-\gamma \|x_a - x_b\|^2)$$

$$Polynomial\colon K(x_a, x_b) = (\gamma \langle x_a, x_b \rangle + r)^d$$

where $\gamma$, $r$ and $d$ are kernel specific hyperparameters. The separating hyperplane is linear in higher dimensional space, but translates to a non-linear decision boundary in the original input space.

*5.4.7.4. Practical notes on the SVM*

Training an SVM requires tuning the hyperparameters, namely the cost parameter *C* and the kernel function. In addition, some kernels have kernel-specific hyperparameters, for example the degree of a polynomial kernel. For an SVM with linear kernel, the model parameters $\beta_1, \ldots, \beta_p$ can be interpreted as feature importance scores. There is no straightforward way to extract importance scores from non-linear SVMs. The SVM algorithm can also be adapted to regression problems, though a discussion of support vector regression machines is beyond the scope of this text.



Input space High dimensional feature space

Figure 20. In the original input space the observations are not linearly separable. However, by transforming to higher dimensional space the red and green classes can be separated by a linear hyperplane.

## 5.5. Deep learning

Neural networks are a class of machine learning algorithms also referred to as deep learning. Although the first neural networks were developed in the 1960s, training neural networks, especially with many layers, was a computationally challenging task given the large number of parameters in these models. The deep learning field has made a lot of progress in the last few years thanks to the increase in computational power and speed achieved by graphics processing units (GPUs) and the availability of very large datasets. Neural networks are highly flexible classification and regression algorithms and have outperformed other machine learning approaches in various research fields (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015) .

### 5.5.1. Perceptron and artificial neurons

The building stones of neural networks are perceptrons and artificial neurons. A perceptron is relatively simple model that takes multiple input variables and returns a single binary output. Each feature has a weight and the weighted sum of the input variables determines

the output. The perceptron returns 1 when the weighted sum exceeds a certain threshold and 0 otherwise. For a perceptron with $p$ input features,

$$\text{perceptron output} = \begin{cases} 0 \text{ if } \sum_{j=1}^{p} w_j x_{ij} + b \ \leq 0 \\ 1 \text{ if } \sum_{j=1}^{p} w_j x_{ij} + b > 0 \end{cases}$$

where $w_j$ is the weight for the $j$th feature and $b$ is the negative of the threshold. $b$ is also called the bias of the perceptron. The terminology used by the deep learning and statistics community is different, but weights correspond to model coefficients, whereas the bias term can be seen as the intercept. We refer to the weighted sum of input features plus the bias term the *input* of the perceptron.

$$z_i = \sum_{j=1}^{p} w_j x_{ij} + b$$

In fact, a perceptron is equivalent to a linear SVM, although both algorithms typically use different optimization techniques (Collobert & Bengio, 2004). The binary nature of the perceptron output is a limitation of this model. Changes in the input features either result in a reversal of the output or no change at all. Artificial neurons are modified perceptrons that apply an activation function to the input $z$ and produce a continuous outcome, allowing for a more nuanced response to changes in the input features. A sigmoid neuron applies the logistic function to $z$

$$f(z) = \frac{1}{1 + \exp(-z)}.$$

The output from a sigmoid neuron is thus a continuous number between 0 and 1, and can be interpreted as the probability of belonging to outcome class 1. Other activation functions include the hyperbolic tangent, linear and rectified linear functions (Table 5). A graphical representation of an artificial neuron helps explain its similarities with a biological

neuron (Fig. 21). The input features can be compared to dendrites, through which information is received, the circle depicting the activation function is similar to the soma or cell body and the output arrow is the equivalent of the axon, which fires when the electrical potential in the cell surpasses a threshold.

Table 5. Activation functions in artificial neurons.

| **Activation function** | **Formula** | **Output range** |
|---|---|---|
| Sigmoid | $f(z) = \frac{1}{1+\exp(-z)}$ | $[0, 1]$ |
| Hyperbolic tangent | $f(z) = \tanh(z)$ | $[-1, 1]$ |
| Linear | $f(z) = z$ | $]-\infty, \infty[$ |
| Rectified linear | $f(z) = \max(0, z)$ | $[0, \infty[$ |
| Softmax (for $k$ outcome classes) | $f(z)_j = \frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)}$ | $[0, 1]$ and $\sum_{k=1}^{K} f(z)_k = 1$ |

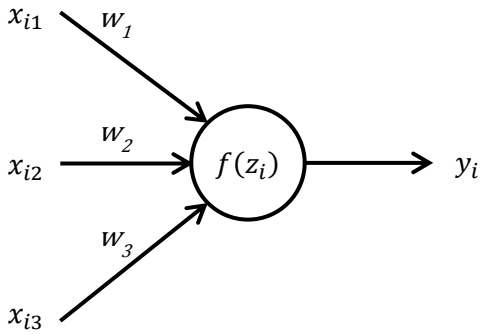



Figure 21. Artificial neuron with three input features, applying activation function $f(z)$ to the weighted sum of features $z_i$ to produce outcome $y_i$.

### 5.5.2. Neural network

The output of a neuron can be used as input feature to another neuron and several neurons can be connected to form a network. The result is a highly flexible non-linear prediction model. In the graphical representation of a neural network, the neurons are

arranged in layers (Fig. 22). The input features are drawn as circles, which form the input layer of the network. In the next layers, circles represent artificial neurons. The layers of neurons from which the output serves as input for the subsequent layer of neurons are called hidden layers. The number of hidden layers and the number of neurons in each one are hyperparameters of the neural network. Finally, the last layer is called the output layer. For a binary outcome, the output layer consists of a single sigmoid neuron, predicting the probability of outcome class 1. If the outcome is categorical with $K$ categories, a layer of $K$ softmax neurons can be used (Table 5). The $k^{th}$ softmax neuron gives the probability of the outcome belonging in the $k^{th}$ category. A continuous outcome is most suitably modelled by a single linear output neuron. Each connecting arrow in the neural network graph has a weight attached to it and each neuron is associated with a bias term. It is obvious that the number of parameters in a neural network rapidly grows with the complexity of the model and that fitting a neural network can be computationally expensive.

With the exception of a single perceptron or artificial neuron it is not straightforward to interpret the weights and biases and derive feature importance scores from a neural network. This limitation implies that these algorithms are not suitable for analysis problems where understanding how the predictor variables interact is important. On the other hand, neural networks are very flexible classification and regression models and often outperform other machine learning approaches for the modelling of large datasets.

The weights and biases of a neural network are iteratively updated during the model training process using mini-batch gradient descent, an iterative optimization algorithm. To compute the gradient of the cost function, backpropagation is used, which is an efficient algorithm that backwardly propagates the error through the neural net. Each gradient descent iteration involves a forward pass through the network to compute the cost function and a backward pass to compute the derivative of the cost function and update the model parameters.

Whereas other machine learning algorithms are usually trained using cross-validation, the neural network community has a preference for the training-validation-test data approach, whereby algorithms are fit on the training data, then the model that achieves the best prediction in the validation data is selected and in a final step that model is assessed on the test data. This approach avoids the computationally expensive part of cross-validation, namely the model fitting, but requires larger sample sizes as the data are split in three independent subsets.

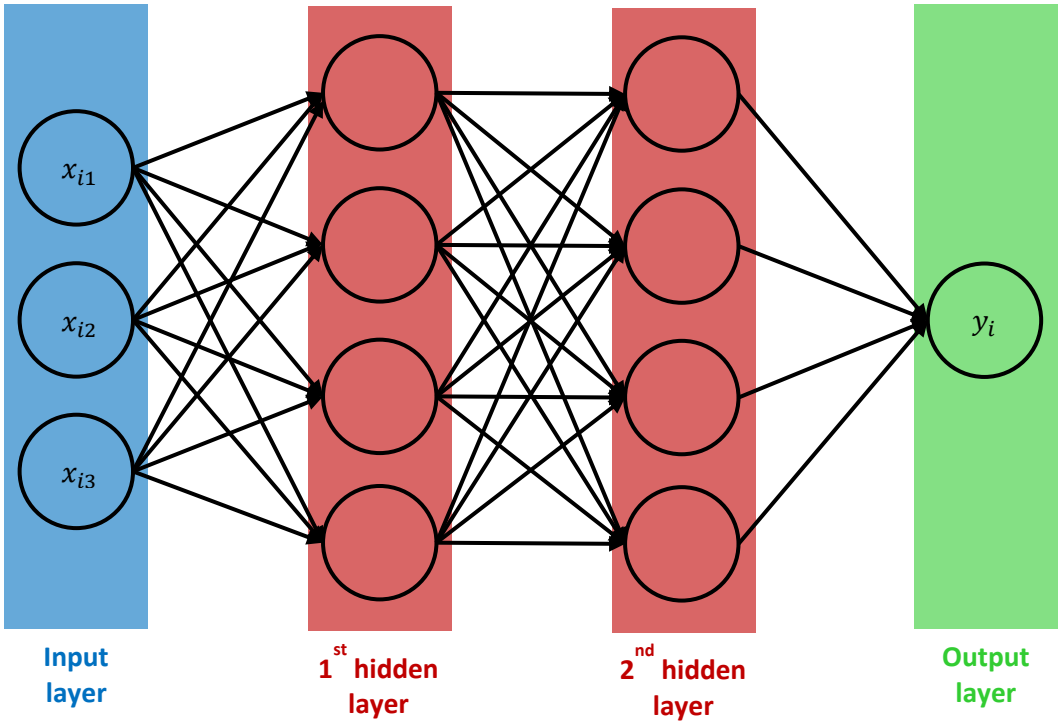



Figure 22. A neural network with three input features, two hidden layers of four hidden neurons each and a single output neuron.

With the large number of model coefficients, neural networks can easily overfit to the training data. One way to prevent overfitting is early stopping; the model training is stopped as soon as the prediction accuracy in the validation data does not improve further. A second strategy is applying drop-out during model fitting, where a random selection of input features and neurons is dropped for the duration of a single training iteration. By

iteratively excluding parts of the neural network, the model is forced to learn more robust patterns (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). Another alternative is to add a regularization penalty to the cost-function which is minimized.

The design of a neural network, i.e. defining the structure of the network, including the number of hidden layers, the number of hidden neurons and the connections between the neurons, choosing the activation functions and deciding the gradient descent parameters (number of iterations, mini-batch size, learning rate) is not a straightforward task. There is no consensus in the field of the preferred strategy to optimize hyperparameters, though a pragmatic approach is to build a relatively small network that achieves some prediction and then try to improve its performance by repeatedly tuning the hyperparameters one by one (Nielsen, 2015). Alternatively, a grid search can be performed but this involves many repetitions of fitting a computationally intensive model. A random search, where hyperparameter values are randomly sampled provides a faster way to explore hyperparameter settings (Bengio, 2012).

The neural network described in the preceding paragraphs is sometimes referred to as a feedforward multi-layer perceptron (MLP), even though the neurons need not be perceptrons and are usually sigmoid neurons. In the MLP each layer is fully connected to the next one. More complex architectures of neural networks exist, which take the nature of the input data into account. For example, convolutional networks are well suited for image recognition (Krizhevsky, Sutskever, & Hinton, 2012). Images are translated into matrices of pixel intensities. Convolutional networks have hidden layers which are organized in parallel and each layer scans the input image to detect a visual feature. The network learns the weights of the hidden layers from the data and thus automatically learns visual features that are useful for classification. The architecture of convolutional networks reduces the number of coefficients compared to a fully connected MLP and makes these algorithms faster to train. A second example is recurrent neural networks.

These networks have cyclic connections between neurons and are ideal for the analysis of sequential data. Given a sequence of input data, a recurrent neural network predicts the next item. Recurrent neural networks have been successful in speech recognition (Graves, Mohamed, & Hinton, 2013).

## 5.6. Summary

Traditional statistical methods like linear and logistic regression are limited in the number of variables a single model can fit and struggle when predictor variables are correlated. By consequence, large genetic datasets are analysed in a univariate way and a subsequent correction for multiple testing is applied. In contrast, machine learning and deep learning algorithms can model a large number of correlated predictor variables simultaneously. These methods optimize prediction precision and can be used to build multivariable prediction models for continuous or categorical outcomes. We applied the methods described in this chapter to genetic and gene expression studies.

# 6. Diagnostic classification using machine learning and deep learning applied to brain gene expression data

## 6.1. Introduction

Schizophrenia is a psychiatric complex disorder affecting nearly 1% of the population. It is a highly heritable condition with heritability estimates up to 80% (Gejman, Sanders, & Duan, 2010). GWAS has proven a successful approach for discovering genetic variants that are associated with schizophrenia risk. In the largest GWAS on schizophrenia to date, conducted by the Psychiatric Genetics Consortium (PGC), 128 SNPs in 108 independent genetic loci were identified (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Many of the implicated associations mapped to genes expressed in the brain and to genes involved in the immune response.

It is often not straightforward to derive the molecular effects of a SNP, particularly for variants that lie outside protein coding regions. In addition to the genetic code, gene expression levels can be studied to investigate differences between cases and controls. Various techniques can be used to measure expression levels, including RNA-seq, a next-generation sequencing technique covering the whole transcriptome. Whereas the genetic code in DNA is constant, gene expression levels change over time and are tissue specific. RNA-seq thus takes a snapshot of gene expression in a tissue at a specific point in time. Furthermore, differences in gene expression may be causal to the disease studied or on the contrary may be the consequence of the disease.

Differential gene expression studies compare the expression of each gene between cases and controls and correct the statistical significance threshold for multiple testing. Several genes have been shown to be differently expressed in post-mortem brains of schizophrenia

cases and controls (Bray, 2008; Xin Li & Teng, 2015). Genes related to immune response were upregulated in the hippocampus and dorsolateral prefrontal cortex (DLPFC) of schizophrenia patients (Fillman et al., 2013; Hwang et al., 2013). Differentially expressed genes have also been reported in other tissues, including fibroblasts and peripheral blood cells of schizophrenia patients (Cattane et al., 2015; Sainz et al., 2013).

Machine learning approaches enable large multivariable analyses and can thus be used to simultaneously examine the expression of many genes in a single model. Moreover, these algorithms optimize the accuracy of predicting the outcome of interest, for example diagnostic class. The aim of this study was to classify schizophrenia cases and controls using RNA-seq gene expression data from DLPFC brain tissue. We applied machine learning and deep learning classification algorithms using brain gene expression scores as predictor variables. Three different algorithms, SVM, random forest and neural networks, were trained and their predictive accuracy on test data was compared.

## 6.2. Methods

### 6.2.1. Data

The BrainSeq consortium is a collaboration between the Lieber Institute for Brain Development (LIBD) and seven pharmaceutical industry partners, including Eli Lilly and Company, investigating the molecular mechanisms through which genetic variants associated with psychiatric disorders act (Lieber Institute for Brain Development, 2017; Schubert et al., 2015). In addition to genetic data, the BrainSeq project also generates and analyses epigenetic and transcriptomic datasets from different brain regions and will make them publicly available after completion of the primary analyses. Differential RNA-seq expression analysis of the DLPFC brain region has recently been completed and the results are currently submitted for publication.

Through Eli Lilly, a founding member of BrainSeq, we gained access to the DLFPC gene expression dataset prior to public release. RNA-seq was performed on RNA extracted from the DLPFC of post-mortem human brain tissue donated to the LIBD. The raw sequencing reads (FASTQ files) generated by LIBD were processed using a proprietary pipeline developed and used at Eli Lilly. The RNA-seq pipeline uses open-access and bespoke software and checks for data integrity and potential errors. The raw sequencing reads were put through pre-alignment QC, aligned against human reference genome build 19 and subjected to post-alignment QC (Fig. 23). Next, the expression counts were summarized at the gene level and normalised using log normalisation. Further quality checks included the generation of summary statistics for each normalised sample and a visual inspection of the distributions of all gene expression scores in the data, the variability of the median values and skewness of the samples. The resulting gene expression scores used in our analysis are continuous variables without missing values, each variable corresponding to one gene. In addition to gene expression scores, demographics and sample quality related measures were available (Table 6). There were 408 samples from individuals with schizophrenia and healthy controls in the LIBD dataset.

### 6.2.2. Statistical methods

#### *6.2.2.1. Data splitting and pre-processing*

Three different machine learning and deep learning algorithms were applied to predict schizophrenia case/control status from brain gene expression data. The predictive accuracy of a classification algorithm, i.e. the proportion of subjects that are correctly classified, is assessed on a test dataset independent from the data used to train the model. Therefore, the LIBD dataset was randomly split in training (80%) and test (20%) subsets. The training data were used to build the machine learning models, which were then applied to the independent test data to measure their predictive performance.

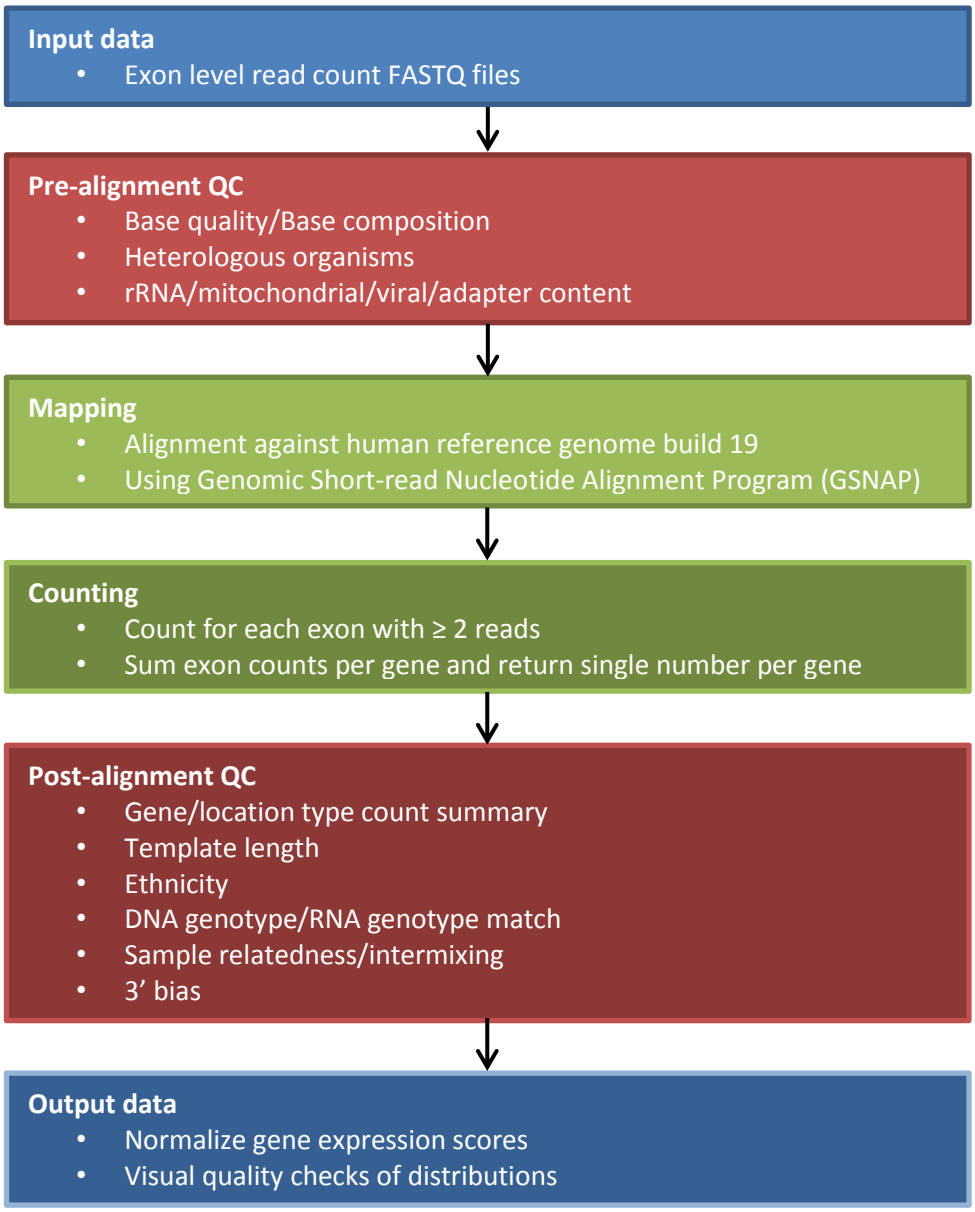



Figure 23. Eli Lilly RNA-seq quality control pipeline. QC: quality control.

The gene expression scores were adjusted for demographic and sample related variables to control for confounding covariates (Table 6). The effect of the covariates on gene expression was estimated using linear regression in the training set control samples. By only modelling the covariates in the controls we separate the covariate effects from differences caused by the disease status (Dukart, Schroeter, Mueller, & Alzheimer's Disease Neuroimaging Initiative, 2011). Subsequently, these regression models were used to correct the gene expression scores in the training and the test data.

Table 6. Demographics and sample quality measures controlled for in gene expression analyses.

| Covariate | Description |
|---|---|
| Age | Age in years |
| Sex | Sex of the patient |
| Ethnicity | Ethnicity of the patient |
| Smoking status | Whether the patient consumed tobacco. When this information was not available it was derived from nicotine and cotinine toxicology analysis. |
| pH | Acidity of cerebellar tissue, a measure relating to basic tissue quality. |
| RNA integrity number (RIN) | RNA quality measure capturing RNA degradation. Ranges from 1 (degraded) to 10 (intact). |
| Post mortem interval (PMI) | Time in hours between death of the patient and freezing of the brain. RNA degrades over time thus shorter PMI is preferable. |

*6.2.2.2. Machine learning algorithms*

SVM, random forest and neural network algorithms were trained to classify schizophrenia cases from controls (Fig. 24). The SVM algorithm has previously been successfully applied to gene expression data in an antidepressant response study and therefore we chose this machine learning classifier for our study (Malki et al., 2017). We compared the SVM to random forest, another well-established classifier for high dimensional data. Neural networks are highly flexible classification algorithms which have outperformed other machine learning methods in various fields. Hence, we explored the use of this deep learning method on gene expression data. The hyperparameters of the algorithms were optimized using five-fold cross-validation.

An SVM separates the outcome classes by defining a decision boundary and maximizing the distance between the data and that boundary (for details of the SVM algorithm, see section 5.4.7. on page 86). Different kernels can be applied to project the input variables to higher dimensional space and compute non-linear decision boundaries. The cost-parameter of an SVM determines how strictly the separation between the classes is enforced. Five-fold cross-validation was used to tune the kernel type, kernel specific parameters and the value of the cost-parameter (Table 7).

Table 7. Hyperparameters of machine learning algorithms.

| **Algorithm** | **Hyperparameter** | **Values tested in tuning process** |
|---|---|---|
| SVM | Kernel type | Linear, radial, polynomial |
| | Cost | 0.01, 1, 100 |
| | Degree of polynomial kernel (kernel specific parameter) | 2, 3, 4 |
| | $\gamma$ (kernel specific parameter) | $10^{-4}$, $10^{-5}$, $10^{-6}$ |
| Random forest | $r$ (number of predictors randomly selected for tree building) | $\log(p)$, $\sqrt{p}$, $p$ |
| Neural network | Number of hidden layers | 1, 2, 3 |
| | Number of hidden neurons per hidden layer | 500, 1,000, 2,000, 5,000, 10,000 |

$p =$ total number of predictors

A random forest is an ensemble of decision trees (for details of the random forest algorithm, see section 5.4.6.2. on page 84). Each tree is constructed using a bootstrapped sample and a random selection of predictor variables. The number of predictors that is randomly selected for the construction of each tree ($r$) was optimized using five-fold cross-validation (Table 7). We built random forests with 1,000 trees. Including more trees does

not induce overfitting of the model to the training data, and it is thus not necessary to tune this hyperparameter.

Lastly, neural networks were applied to solve the classification problem (for details of the neural network algorithm, see section 5.5.2. on page 93). These are highly flexible algorithms with a large number of parameters. As the computational cost of neural networks increases rapidly with the complexity of the design, a pragmatic approach was used for hyperparameter tuning. Instead of defining a list of hyperparameter values to be evaluated prior to cross-validation, we started by training a simple neural network with 1 hidden layer of 500 neurons. We then then increased the complexity of the model by varying the number of hidden layers and hidden neurons whilst taking guidance from the cross-validated prediction accuracy (Table 7). The parameters of backpropagation, the iterative optimization method used to estimate neural network parameters, were fine tuned for each complexity level of the algorithm. To prevent overfitting we applied drop-out to the input (10%) as well as hidden layers (30%), which means that in each training iteration a randomly selected proportion of neurons is excluded.

*6.2.2.3. Variable selection*

Big datasets may contain a large proportion of variables irrelevant to predicting the outcome of interest. Different variable selection techniques can be used to reduce the complexity of the data and improve the signal-to-noise ratio. Including variable selection in the model building process leads to more sparse models and may improve prediction accuracy (Guyon & Elisseeff, 2003).

Initially, machine learning algorithms including the full set of gene expression predictors were built to tune the hyperparameters of the model. After the values of the hyperparameters had been fixed, four different variable selection methods were applied to SVM and random forest models and the results were compared. Three filter methods were

used, where the predictor variables are ranked using t-tests (parametric), Mann-Whitney U tests (non-parametric) or information gain (Kraskov, Stögbauer, & Grassberger, 2004). The top 10, 25, 50, 100, 250, 500, 750, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,500, 10,000, 12,500, 15,000 and 20,000 genes were taken forward as predictors in the machine learning algorithms. The fourth variable selection technique was recursive feature elimination (RFE). This method iteratively excludes the variables that have the smallest impact on the algorithm's prediction accuracy. RFE was used to build SVM and random forest algorithms with 10, 25, 50, 100, 250, 500, 750, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,500, 10,000, 12,500, 15,000 and 20,000 genes as predictors.

In addition, we selected genes based on previously published knowledge. The PGC Schizophrenia GWAS detected 108 significant loci and mapped these to genes (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). We used the list of mapped genes from the PGC study to select genes for inclusion as predictor variables in SVM and random forest models. The PGC mapped the GWAS results to 351 protein coding genes, of which 322 corresponded to genes in our study. Genes were matched through their Entrez GeneID (Maglott, Ostell, Pruitt, & Tatusova, 2011).

No variable selection was applied prior to neural network modelling. As neural networks take interactions between predictors into account, it is less meaningful to exclude predictors on an individual basis. Moreover, historically neural networks were built without variable selection.

*6.2.2.4. Prediction accuracy in independent test set*

For each machine learning approach, the model containing all genes and the model that achieved the highest cross-validated prediction accuracy after variable selection were taken forward. These algorithms were applied to classify schizophrenia case/control status is the LIBD test set and the prediction accuracy was calculated.

### 6.2.3. Software

All analyses were performed in Python 2.7.12. The SVM and random forest algorithms were built using the *scikit-learn* Python library (Pedregosa et al., 2011). The neural net analyses were carried out using the *theano* and *lasagne* Python libraries (Bergstra et al., 2010; Dieleman et al., 2015).

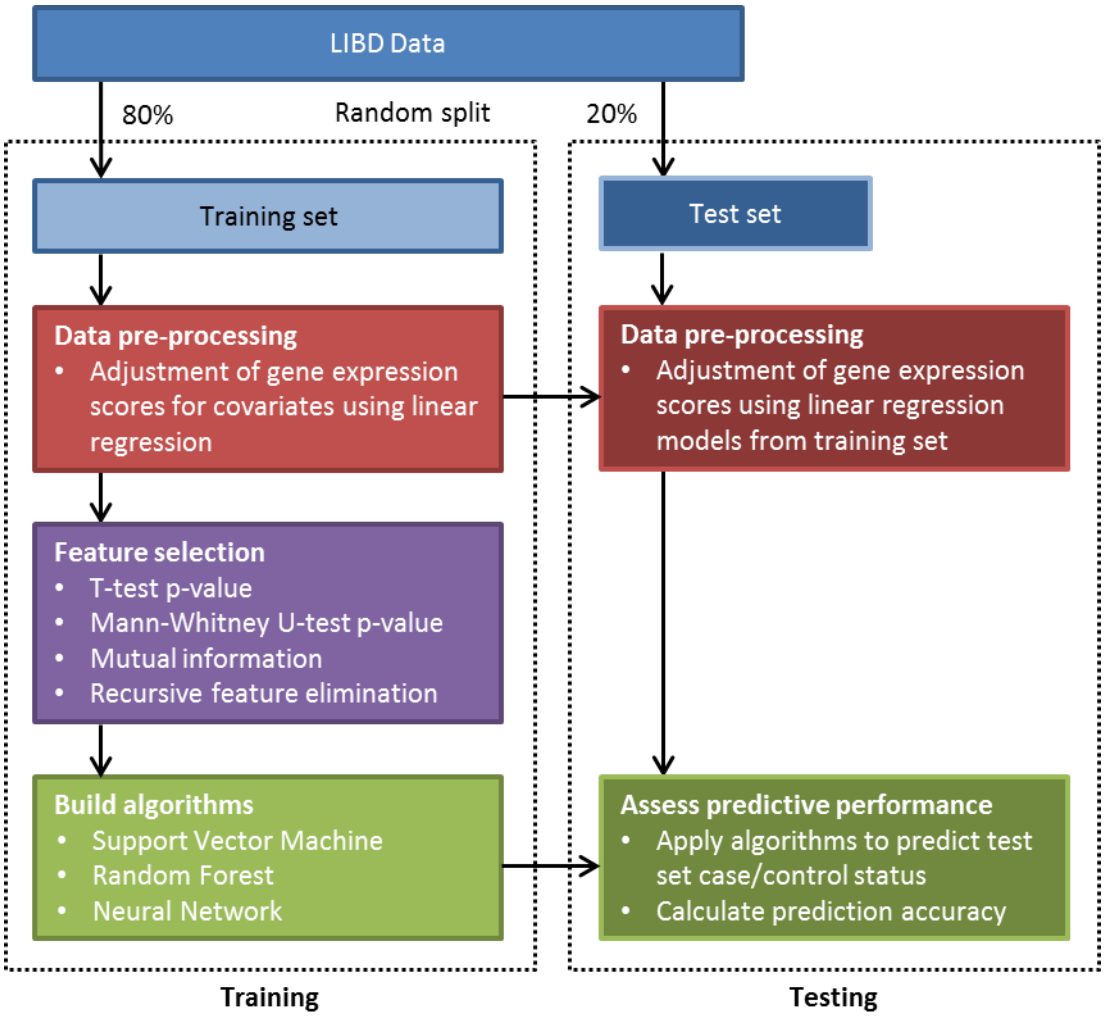



Figure 24. Analysis plan for building machine learning classification models for schizophrenia case/control status.

## 6.3. Results

### 6.3.1. Data description

The LIBD gene expression dataset contained DLPFC brain gene expression scores for 24,383 genes on 408 samples. For one individual the smoking status was missing and could not be estimated, and this sample was excluded from the analysis. No other covariates were missing. There were more controls (55.0%) than schizophrenia cases (45.0%) in the dataset. Most samples were male (67.8%) and the age of the subjects in the dataset ranged between 17 and 97 years (Table 8). The data were randomly split in independent training (80%) and test (20%) sets.

Table 8. Description of outcome and covariates in LIBD training and test data.

| Variable | LIBD Training set | LIBD Test set |
|---|---|---|
| Outcome class, count (%) | | |
| Case | 149 (45.8) | 34 (41.5) |
| Control | 176 (54.2) | 48 (58.5) |
| Sex, count (%) | | |
| Male | 226 (69.5) | 50 (61.0) |
| Female | 99 (30.5) | 32 (39.0) |
| Smoking status, count (%) | | |
| Yes | 159 (48.9) | 36 (43.9) |
| No | 166 (51.1) | 46 (56.1) |
| Ethnicity, count (%) | | |
| Caucasian | 159 (48.9) | 34 (41.5) |
| African American | 151 (46.5) | 43 (52.4) |
| Other | 15 (4.6) | 5 (6.1) |
| Age in years, mean (s.d.) | 45.8 (15.3) | 48.7 (16.2) |
| PMI in hours, mean (s.d.) | 33.8 (18.2) | 34.1 (24.4) |
| pH, mean (s.d.) | 6.5 (0.3) | 6.5 (0.3) |
| RIN, mean (s.d.) | 8.2 (0.8) | 8.1 (1.0) |
| Sample size | 325 | 82 |

### 6.3.2. Machine learning prediction algorithms

Three different machine learning approaches were applied to classify schizophrenia cases from controls using brain gene expression data. After the hyperparameters of the SVM and random forest were tuned, different variable selection techniques were applied to reduce the number if inputs. No variable selection was performed on neural networks.

#### *6.3.2.1. Support vector machine*

The hyperparameters of the SVM were tuned in a model including the full set of genes as predictor variables. The linear kernel reached 83.38% accuracy in the cross-validation, which was higher than the polynomial or radial kernel and was thus carried forward (Table 9). The value of the cost-parameter had little influence and was set to the default value of one.

Next, different variable selection methods were applied to the linear SVM (Fig. 25). Recursive feature elimination retained slightly higher cross-validation accuracy in the training data than t-test and mutual information based ranking. The method using Mann-Whitney U-tests to rank and select variables performed substantially worse than the other methods when lower numbers of variables were selected.

The best predicting model in training set cross-validation was the SVM with 10,000 genes selected by RFE, which predicted with 84.92% accuracy. This is marginally higher than the accuracy achieved by the full set of genes (83.38%). Similar levels of accuracy are obtained by more sparse models, for example an SVM including 250 genes selected by RFE reached 80.31% prediction accuracy. The PGC list of 322 genes achieved 71.08% accuracy.

Table 9. Cross-validated prediction accuracy of SVM hyperparameter values, ordered by decreasing cross-validated prediction accuracy.

| **Kernel** | **Cost parameter** | **$\gamma$-parameter** | **Degree of polynomial kernel** | **5-fold cross-validated prediction accuracy (%)** |
|---|---|---|---|---|
| Linear | 1 | - | - | 83.38 |
| Linear | 0.01 | - | - | 83.38 |
| Linear | 100 | - | - | 83.38 |
| Radial | 0.01 | - | $10^{-5}$ | 80.92 |
| Polynomial | 1 | 2 | $10^{-5}$ | 80.00 |
| Polynomial | 1 | 3 | $10^{-4}$ | 73.85 |
| Radial | 1 | - | $10^{-4}$ | 70.77 |
| Polynomial | 0.01 | 2 | $10^{-5}$ | 70.46 |
| Polynomial | 0.01 | 3 | $10^{-5}$ | 65.23 |
| Polynomial | 100 | 3 | $10^{-5}$ | 64.62 |
| Radial | 0.01 | - | $10^{-6}$ | 63.38 |
| Radial | 100 | - | $10^{-5}$ | 62.46 |
| Polynomial | 0.01 | 2 | $10^{-4}$ | 62.46 |
| Polynomial | 0.01 | 4 | $10^{-5}$ | 58.15 |
| Polynomial | 1 | 3 | $10^{-5}$ | 57.85 |
| Polynomial | 1 | 3 | $10^{-6}$ | 56.62 |
| Polynomial | 0.01 | 3 | $10^{-4}$ | 56.62 |
| Polynomial | 100 | 3 | $10^{-4}$ | 55.08 |
| Polynomial | 0.01 | 4 | $10^{-4}$ | 55.08 |
| Polynomial | 1 | 4 | $10^{-6}$ | 55.08 |
| Polynomial | 1 | 2 | $10^{-4}$ | 54.15 |
| Polynomial | 100 | 2 | $10^{-4}$ | 54.15 |
| Radial | 100 | - | $10^{-4}$ | 54.15 |
| Polynomial | 1 | 2 | $10^{-6}$ | 54.15 |
| Polynomial | 1 | 4 | $10^{-5}$ | 54.15 |
| Polynomial | 100 | 2 | $10^{-6}$ | 54.15 |
| Polynomial | 0.01 | 3 | $10^{-6}$ | 54.15 |

Table 9 (continued). Cross-validated prediction accuracy of SVM hyperparameter values, ordered by decreasing cross-validated prediction accuracy.

| Kernel | Cost parameter | $\gamma$-parameter | Degree of polynomial kernel | 5-fold cross-validated prediction accuracy (%) |
|---|---|---|---|---|
| Radial | 1 | - | $10^{-6}$ | 54.15 |
| Polynomial | 100 | 3 | $10^{-6}$ | 54.15 |
| Radial | 100 | - | $10^{-6}$ | 54.15 |
| Radial | 1 | - | $10^{-5}$ | 54.15 |
| Polynomial | 100 | 2 | $10^{-5}$ | 54.15 |
| Polynomial | 100 | 4 | $10^{-4}$ | 54.15 |
| Polynomial | 1 | 4 | $10^{-4}$ | 54.15 |
| Polynomial | 100 | 4 | $10^{-5}$ | 54.15 |
| Radial | 0.01 | - | $10^{-4}$ | 54.15 |
| Polynomial | 0.01 | 2 | $10^{-6}$ | 54.15 |
| Polynomial | 0.01 | 4 | $10^{-6}$ | 54.15 |
| Polynomial | 100 | 4 | $10^{-6}$ | 54.15 |

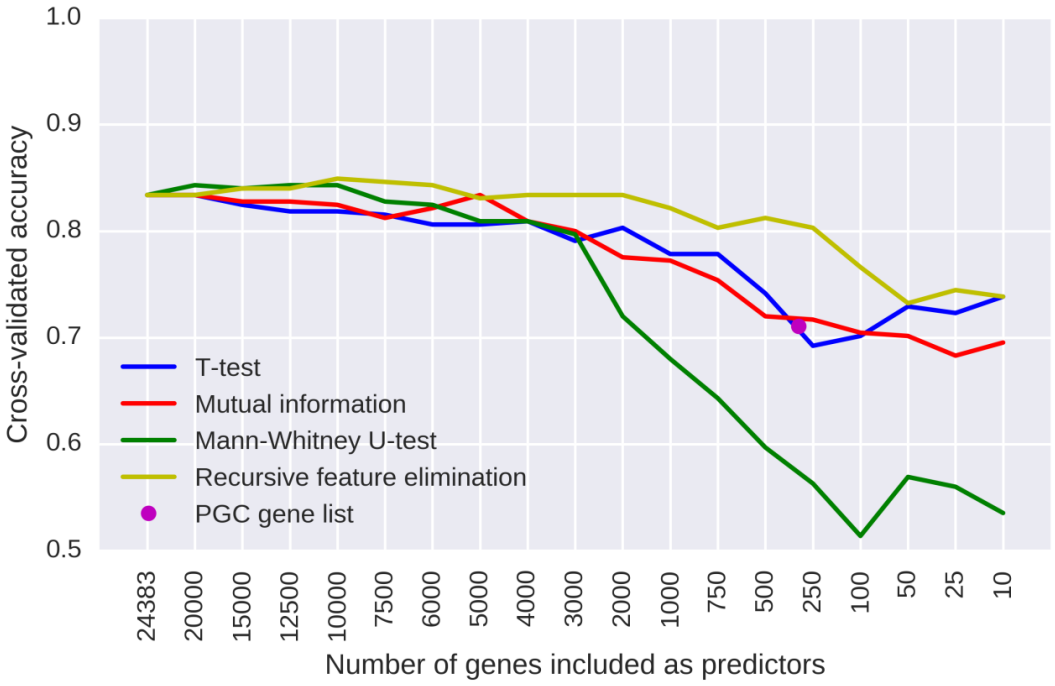



Figure 25. Cross-validated prediction accuracy for SVM using different variable selection methods.

*6.3.2.2. Random forest*

Overall, the prediction accuracy obtained from random forest algorithms was lower than from SVMs. The only hyperparameter of the random forest, $r$ (the number of randomly selected predictors for construction of the individual trees), was cross-validated and set to the total number of predictors (24,383 genes) (Table 10). Allowing each tree to choose from the full set of genes thus performed better in cross-validation than selecting a random subset of predictors. This method is a special case of random forest also called bagging. With $r = p$, the cross-validated prediction accuracy was 75.38%.

Table 10. Cross-validated prediction accuracy of random forest hyperparameter values, ordered by decreasing cross-validated prediction accuracy.

| $r$ **parameter** | **5-fold cross-validated prediction accuracy (%)** |
|---|---|
| $p$ | 75.38 |
| $\sqrt{p}$ | 72.92 |
| $\log(p)$ | 69.54 |

$p$: total number of predictor variables.

Four different variable selection methods were used to reduce the number of genes in the random forest (Fig. 26). The difference between RFE, t-test based ranking or mutual information based ranking was trivial. However, variable selection using Mann-Whitney U-test based ranking lead to markedly lower accuracy scores. The accuracy of random forest algorithms built using RFE or t-test based ranking ranged between 72.92% and 76.62% and was roughly constant irrespective of the number of predictors included. The model with 322 genes from the PGC GWAS study reached 70.15% accuracy.

Several random forests achieved the maximum cross-validated accuracy of 76.62%: the models with 4,000 and 10,000 genes selected based on mutual information ranking and the

models with 250, 2,000 and 6,000 genes selected by RFE. The random forest with 250 genes selected using RFE was carried forward because this was the sparsest model achieving maximal prediction accuracy.

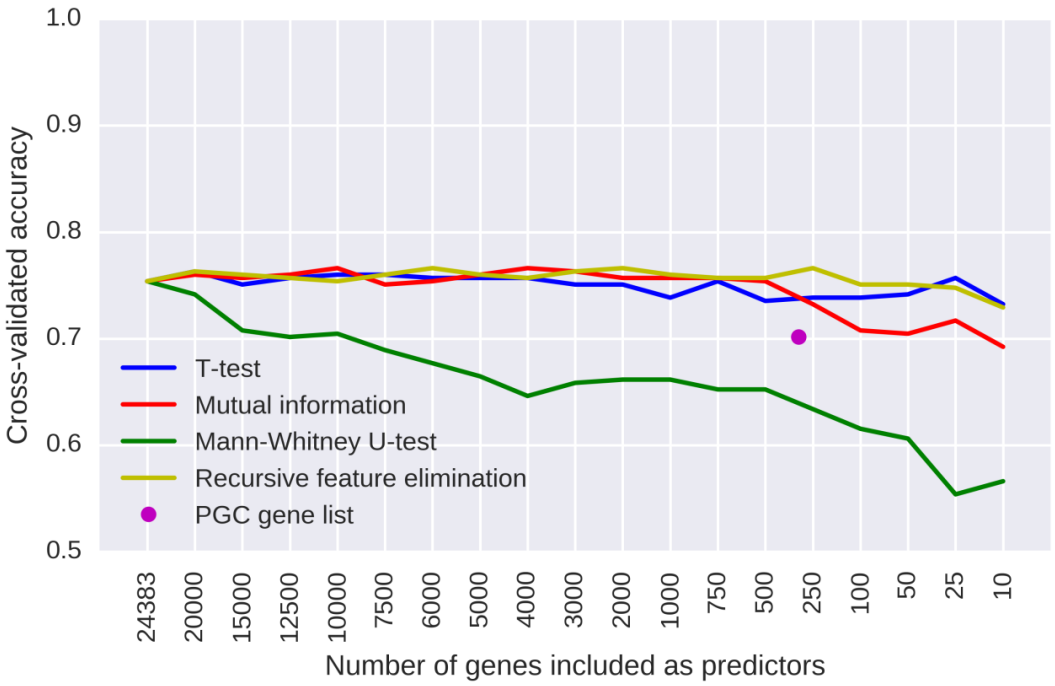



Figure 26. Cross-validated prediction accuracy for random forest using different variable selection methods.

*6.3.2.3. Neural network*

The tuning parameters of the neural network were optimized using five-fold cross-validation. We started with a network with one hidden layer of 500 hidden neurons and expanded the design complexity by increasing the number of neurons and the number of hidden layers. Adding more hidden neurons to a model with a single hidden layer had no notable effect on the cross-validated accuracy (Table 11). When a second hidden layer was added, the accuracy decreased by approximately six percentage points. Again, the number of neurons in the hidden layers had a trivial effect on the accuracy of two-layer neural networks. Increasing the number of hidden layers to three substantially lowered the prediction accuracy, after which we concluded the network design. The neural network that

achieved maximal cross-validated prediction accuracy (79.06%) had 1 hidden layer of 2,000 hidden neurons. Thus, the best prediction accuracy achieved by neural networks was higher than the accuracy of random forests, but lower than that of SVMs. No variable selection was applied to the neural network algorithms; these models were built using all 24,383 genes as predictor variables.

Table 11. Cross-validated prediction accuracy for neural networks with varying design complexity, ordered by decreasing cross-validated prediction accuracy.

| Number of hidden layers | Number of hidden neurons per layer | 5-fold cross-validated prediction accuracy (%) |
|---|---|---|
| 1 | 2,000 | 79.06 |
| 1 | 10,000 | 79.05 |
| 1 | 5,000 | 78.14 |
| 1 | 500 | 77.83 |
| 1 | 1,000 | 77.22 |
| 2 | 5,000 | 72.91 |
| 2 | 1,000 | 72.30 |
| 2 | 10,000 | 72.00 |
| 2 | 500 | 71.98 |
| 2 | 2,000 | 71.98 |
| 3 | 500 | 56.32 |

### 6.3.3. Validation of predictive power in independent test set

Selected algorithms were taken forward and applied to classify schizophrenia cases from controls in independent test data. The prediction accuracy of the SVM, random forest and neural net algorithms including the full set of 24,383 genes as predictor variables, the SVM and random forest models that achieved the highest accuracy after variable selection and the sparse SVM model with 250 genes selected by RFE were assessed in the test dataset. The machine learning algorithms score high prediction accuracy in the test set, ranging

between 69.51% and 78.05% (Table 12). The test set prediction accuracy was however lower than estimated by cross-validation in the training data. The SVM algorithms scored highest in the cross-validation and also ranked highest when applied to the test set. Examining the predictive characteristics of the SVM with 10,000 variables selected through RFE (the highest ranking algorithm in cross-validation) applied to the test set indicates that misclassification is balanced between cases and controls. The sensitivity and specificity of the algorithm were 73.53% and 79.17%, respectively (Table 13).

Table 12. Prediction accuracy of selected algorithms in independent test data, ordered by decreasing cross-validated prediction accuracy.

| | | | Prediction accuracy (%) | |
|---|---|---|---|---|
| **Algorithm** | **Number of genes** | **Variable selection** | **5-fold CV training set** | **Test set** |
| SVM | 10,000 | RFE | 84.92 | 76.83 |
| SVM | 24,383 | None | 83.38 | 78.05 |
| SVM | 250 | RFE | 80.31 | 75.61 |
| Neural network | 24,383 | None | 79.06 | 73.17 |
| Random forest | 250 | RFE | 76.62 | 69.51 |
| Random forest | 24,383 | None | 75.38 | 69.51 |

CV: cross-validated; RFE: recursive feature elimination; SVM: support vector machine.

Table 13. Classification measures of SVM with 10,000 variables selected through RFE applied to test set.

| **Measure** | **Score in test set (%)** |
|---|---|
| Accuracy | 76.83 |
| Sensitivity | 73.53 |
| Specificity | 79.17 |
| PPV | 71.43 |
| NPV | 80.86 |

NPV: negative predictive value; PPV: positive predictive value

### 6.3.4. Summary of results

All algorithms achieved high prediction accuracy in the cross-validated training set and in the independent test set. The SVM outperformed the random forest and neural network in terms of prediction accuracy. Furthermore, RFE yielded slightly better results than the other variable selection methods we applied. The best performing algorithm in cross-validation was an SVM with 10,000 genes selected through RFE, which achieved 76.83% accuracy in the test data.

## 6.4. Discussion

Three different machine learning methods were used to classify schizophrenia cases from controls using gene expression data from the DLPFC brain region. All methods were successful in predicting case/control status, with accuracy scores up to 78.05% in the independent test dataset. The SVM models achieved the highest accuracy in the cross-validation of the training data and also when applied to test data. Neural network was the second most accurate method, followed by random forest, which still classified nearly 70% of test samples correctly.

Overall, the accuracy in the test data is lower than the accuracy estimated by cross-validation in the training data. This is to be expected. Although cross-validation splits the data is separate folds and the accuracy is estimated from predicting how algorithms score on left-out folds, across the entire process each sample is used in both the training in testing folds. Furthermore, cross-validation is performed in order to choose the model parameters that result in the highest accuracy. Precisely because the algorithm is selected based on the highest accuracy, this estimate of the prediction accuracy is likely to be overly optimistic. To obtain an unbiased, generalizable estimate of prediction accuracy, the

algorithm needs to be evaluated on an independent test set and these test data must not be involved in any step of the algorithm training process.

Variable selection is used to reduce the dimensions of the predictor space and to improve prediction accuracy. In this study, four different techniques for variable selection were compared. None of the methods led to marked improvements in accuracy, although sparser models that achieved approximately the same accuracy as the models with 24,383 genes were identified. RFE was the method that performed best when training SVM algorithms. For random forests, RFE, t-test and mutual information based ranking methods achieved similar results. Variable selection based on the non-parametric Mann-Whitney U-test ranking clearly performed worse than the other methods.

RFE variable selection on the SVM algorithm identified a set of 250 genes which achieves approximately the same accuracy as all 24,383 genes. This hints that most predictive power can be summarized in 250 genes. The list of genes can be investigated in more detail using bioinformatics approaches such as gene enrichment analysis and pathway analysis. In this way, genes selected by machine learning may increase our understanding of the biological processes underlying schizophrenia.

Traditionally, RNA-seq data are analysed in differential expression studies, where the expression of genes in cases and controls is compared statistically. Each gene is independently tested and the significance of the difference in expression is assessed. In contrast, we used machine learning to select a set of 250 genes which combined hold most predictive power to classify cases from controls. This list of genes was obtained from an SVM algorithm that models all genes simultaneously. It is important to note these methodological differences when interpreting our results and comparing them to the differential expression literature. The list of genes identified by machine learning and by differential expression analysis may show little or no overlap. The strength of machine

learning lies in its ability to pick up combinations of genes that optimize the prediction of the outcome of interest, although these genes may not have a large effect individually.

In addition to performing variable selection, which is based on information in the training data, we also evaluated SVM and random forest models which included 322 genes derived from the PGC schizophrenia GWAS study as predictors. The cross-validated accuracy of the PGC list SVM was in the range of what can be expected for this number of predictors, based on the results from t-test and mutual information variable selection methods. However, RFE outperformed the PGC gene list algorithm substantially. In the random forest models the differences in predictive performance were smaller, but again variable selection techniques resulted in models with higher accuracy than the PGC list random forest. Nevertheless, this suggests that information from a study of variability in genetic sequence contains some signal that translates to gene expression levels. This is an encouraging finding, as the link between genetic variability and gene expression is not self-evident. A gene expression study looking at 5 genes identified through GWAS (*ZNF804A*, *OPCML*, *RPGRIP1L*, *NRGN* and *TCF4*) found no difference in gene expression levels in brain samples of schizophrenia cases and controls (Umeda-Yano et al., 2014). Moreover, in addition to the type of data studied (DNA variants versus RNA gene expression), there are important methodological differences between the PGC GWAS and our study. In the GWAS SNPs are studied univariately, whereas we used a machine learning approach to model all gene expression scores simultaneously.

Neural networks are very flexible models and have been successfully put to use in many advanced applications such as computer vision and speech recognition. In several fields they have outperformed classic machine learning approaches such as the SVM (LeCun et al., 2015). In our study, neural networks ranked second best in terms of prediction accuracy, after SVM. The limited sample size in our study may be an explanation for this. The training data contained 325 individuals. In comparison, successful applications of neural networks

are often trained on large datasets with thousands or millions of training examples. It is possible that the strengths of the neural network approach might come into play in larger datasets and that with larger samples neural networks may outperform SVM in gene expression studies. Yet, the expectation that larger training datasets will increase the usefulness of the algorithm holds for all machine learning approaches.

In conclusion, gene expression data from the DLPFC region in human brain tissue can classify schizophrenia cases from controls using machine learning algorithms. SVM was the most accurate approach, followed by neural network and random forest. Variable selection was applied to identify a small subset of genes that have the same predictive power as the full set of genes. A set of 322 genes derived from a large schizophrenia GWAS achieved lower accuracy than our best variable selection method, which utilizes the machine learning algorithm itself. Larger sample sizes will further increase the utility of machine learning in gene expression studies.

# 7. Machine learning algorithms for pharmacogenetic prediction in an anti-diabetic clinical trial

## 7.1. Introduction

Type 2 diabetes mellitus (T2DM) is an endocrinologic condition in which blood sugar levels are increased due to insulin resistance and insufficient insulin secretion. In addition to lifestyle changes, T2DM is treated with drugs such as metformin, sulphonylureas, thiazolidinediones and gliptins. A few PGx associations with anti-diabetics have been discovered and replicated to date. Glycated hemoglobin (HbA1c) levels, a measure indicative of the three-month average plasma glucose concentration, show PGx association with several anti-diabetic drugs: *ATM* variants are associated with HbA1c levels in metformin treatment, *TCF7L2*, *KCNJ11* and *CYP2C9* with sulfonylureas, *PPARG* with thiazolidinediones and *CTRB1/2* with gliptins. Furthermore, *SLC22A1* is associated with metformin induced gastrointestinal adverse effects, and *GSTT1* and *CYP2C19* are associated with troglitazone induced hepatotoxicity. *CYP2C9* loss-of-function variants lead to higher sulfonylurea drug levels and are associated with hypoglycaemia. Although these are robust, replicated genetic associations, the effect sizes are not large enough to be of clinical utility as a biomarker to guide treatment (Daniels et al., 2016; Zhou, Pedersen, Dawed, & Pearson, 2016). Discovery of genetic biomarkers with clinical relevance will be very valuable given the increasing prevalence of T2DM (World Health Organization, 2016).

Most PGx studies in anti-diabetics were candidate gene studies, focusing on a small number of genes selected based on prior biological knowledge. Genome-wide studies expand the search for associations across the whole genome and thus take a hypothesis-free perspective, though typically genetic variants are analysed independently. Considering that PGx effects could well be due to a combination of multiple variants with moderate or weak

effect sizes, it is appropriate to perform a multivariable analysis and assess several genetic variants simultaneously. Unlike traditional statistical models, machine learning can handle datasets with large numbers of variables, even if these variables are correlated.

In this study we apply regression trees, random forest and elastic net to an anti-diabetic cross-over clinical trial, with the aim of predicting PGx changes in efficacy and safety measures. The glucagon receptor antagonist LY2409021 (adomeglivant) was developed for the treatment of T2DM by Eli Lilly. Two randomized, double-blind, placebo controlled Phase 2 clinical trials found that LY2409021 significantly lowered glycated haemoglobin (HbA1c) and fasting blood glucose (FBG) levels. These studies also noted moderate increases in alanine aminotransferase (ALT) and aspartate aminotransferase (AST) levels, indicators of hepatic safety, following LY2409021 administration (Kazda et al., 2016). A further Phase 2 trial confirmed the superior efficacy of LY2409021 compared to placebo, and similar efficacy when compared to an active comparator sitagliptin. However, the safety aspect of the study showed that the LY2409021 treated group had increased weight, hepatic fat, hepatic ALT and AST levels and systolic blood pressure (SBP) (Guzman et al., 2017). Finally, a 6-week multicentre, randomized, double-blind, placebo-controlled Phase 2 cross-over trial was conducted to assess the safety and efficacy of LY2409021. This study found that LY2409021 indeed lowered HbA1c and FBG, but increased SBP and aminotransferase levels (ALT and AST) (Kazda et al., 2017). Although LY2409021 lowers glucose related measures, there are some safety concerns that may impact its usefulness for chronic antidiabetic therapy and Eli Lilly discontinued the development of this drug (Adis International Ltd, 2017).

## 7.2. Methods

### 7.2.1. Data

Through our collaboration with Eli Lilly we received permission to analyse clinical and genetic data from the cross-over clinical trial (ClinicalTrials.gov Identifier: NCT02091362) in a machine learning analysis (Kazda et al., 2017). Trial participants received a daily 20mg dose of LY2409021 or placebo for six weeks, followed by a four week washout period and a second six week period of the other treatment (Fig. 27). Our analysis included patients who received at least one dose of both treatments and were genotyped on an Affymetrix-Axiom platform. The primary outcome of the trial was change from baseline to six weeks in mean 24-hour SBP and we extended our study to the secondary outcomes of change HbA1c, FBG, ALT and AST. In addition to clinical descriptives, the data included genotypes of 745,401 SNPs, which we converted to minor allele counts. Our aim was to predict the absolute difference in these five outcome measures after six weeks of LY2409021 and six weeks of placebo treatment using SNPs and clinical baseline variables.

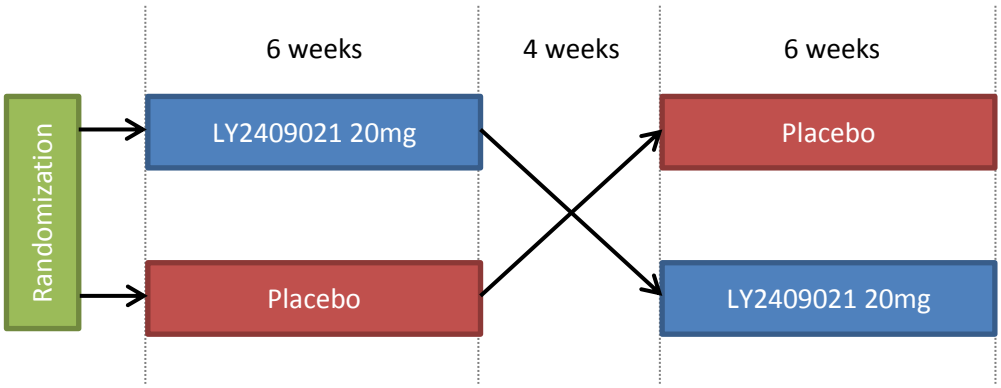



Figure 27. Cross-over trial design comparing LY2409021 and placebo.

### 7.2.2. Statistical methods

Three different machine learning methods were applied to predict PGx differences in SBP, HbA1c, FBG, ALT and AST levels: decision trees, random forest and elastic net. The predictive power of these algorithms was evaluated in an independent test set (Fig. 28).

#### *7.2.2.1. Data splitting and predictive performance assessment*

To obtain an unbiased estimate of the predictive performance of an algorithm, it must be assessed on a dataset independent of the data used to build the model. Therefore, the data were randomly split into separate training (75%) and test (25%) subsets. The training set was used to build the prediction algorithms and subsequently the test set was used to measure the predictive performance on independent data.

The $R^2$ statistic,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

with $y_i$ the observed outcome for subject $i$, $\hat{y}_i$the estimated outcome for subject $i$ and $\bar{y}$ the average of the observed outcomes, measures the proportion of variation in the outcome explained by the algorithm and was used to assess predictive performance. The machine learning algorithms were applied to the test subset to predict the outcome variables and the test set $R^2$ was calculated. In addition, to determine if the models fit the training data well, the $R^2$ on the prediction of the training set outcomes was computed. If an algorithm fails to accurately predict the data it was built with, this indicates poor predictive power and the model is unlikely to perform well on independent test data.

#### *7.2.2.2. Quality control and data pre-processing*

Quality control (QC) is an essential step prior to the analysis of genetic data to ensure validity. To make sure that the test data are completely independent from the model

building process, QC was performed on the training data only and extrapolated to the test data. In the training set, the following QC steps were carried out for SNPs and individuals (Weale, 2010):

- Individual missingness: exclude individuals with call rate < 95%.
- Sex checks: exclude individuals with mismatch between reported sex and genotype sex, using the inbreeding coefficient, *F*.
- Cryptic relatedness: exclude the individual with lowest call rate from pairs that appear related based on the identity-by-descent (IBD) statistic.
- Population stratification: exclude individuals who are outliers in principal components analysis (PCA).
- Heterozygosity and inbreeding: exclude individuals with outlying heterozygosity *F* statistic.
- SNP missingness: exclude SNPs with call rate < 95%.
- Minor Allele Frequency (MAF): exclude SNPs with MAF < 2%. We applied a lenient threshold as the generally used 5% MAF threshold would exclude half of SNPs in this sample.

To perform QC in the test subset, the patients in the test set were scored on the PCA of the training data. Given these factor loadings, the individual level QC steps can be carried out in the test data independently. The SNPs that passed QC in the training sample were retained in the test set and no additional SNP-level QC steps were carried out. Hardy-Weinberg equilibrium tests were not used to exclude SNPs since all study participants suffer from T2DM and association with disease can induce deviations from Hardy-Weinberg equilibrium (Anderson, 2011).

Furthermore, to control for population stratification, the first four principal components were regressed out of the outcome variables and the adjusted outcomes were used as

target variables in all analyses. Principal components were not therefore included in any subsequent models for genetic association of the target variables.

The elastic net and random forest algorithms require complete data. Hence, missing clinical variables were imputed using k-nearest neighbours (KNN) and mean imputation and the imputed datasets were compared by fitting random forest and elastic net models with clinical predictors only. As there was no evident difference between the two imputation methods, the KNN imputed clinical variables were taken forward to the analyses combining clinical and genetic predictors. Missing genotypes were imputed based on linkage disequilibrium (LD) information in PLINK (Purcell et al., 2007). When insufficient LD information is present, this method reduces to imputation with the most common genotype. Our analysis was based on genotyped SNPs and imputation to large reference panels was not performed. As imputation is based on the correlation structure between SNPs, imputed SNPs can introduce a systematic pattern of bias. Multivariable machine learning algorithms can latch on to the imputation pattern, resulting in overfitting and inflated prediction accuracy. Furthermore, adding more features does not necessarily increase the predictive ability of a machine learning model.

*7.2.2.3. Machine learning algorithms*

The three machine learning algorithms (decision trees, random forest and elastic net) were built on the training data to predict changes in the five outcome variables and subsequently the prediction accuracy was evaluated in the independent test data (Fig. 28).

Decision trees are non-linear machine learning models (for details of decision trees, see section 5.4.6.1. on page 81). The Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) algorithm is an adaptation of the classic decision tree algorithm, developed by a collaboration between the University of Wisconsin-Madison and Eli Lilly, with some specific advantages. Firstly, GUIDE performs unbiased variable selection,

meaning that it does not favour splitting on variables with many levels, unlike other decision tree algorithms (Loh, 2002, 2009). In addition, GUIDE has the ability to detect local interactions between predictor variables. The algorithm can be applied to classification as well as regression trees. As Eli Lilly had previous experience with GUIDE and was interested in exploring the use of this method, we applied this machine learning algorithm to our dataset. We constructed GUIDE regression trees to predict changes in the outcome variables and fitted constant (intercept only), simple regression and stepwise multiple regression models in the nodes. The residuals from the node specific regression models are used to split the data in each node into two branches. To prevent overfitting of the models to the training data, the trees were pruned by 10-fold cross validation. The GUIDE algorithm handles missing observations in the predictor variables automatically by imputing with the node means when fitting regression models to the nodes. A limitation of the software used for fitting these trees is that it cannot manage large genome-wide datasets, so the number of genetic predictors was restricted to 1,000 for these analyses.

Random forests are ensembles of decision trees and often outperform single tree algorithms, hence we compared the performance of GUIDE decision trees and random forests (for details of the random forest algorithm, see section 5.4.6.2. on page 84). We built random forests with 1,000 trees, randomly selecting 1/3$^{rd}$ of predictors for inclusion in each tree. The individual trees were constructed using a bootstrapped sample of the data and were not pruned. The random forest algorithm requires complete data, thus missing predictor variables were imputed as described above.

Finally, elastic net models were applied to build linear prediction algorithms and offer an alternative to the non-linear tree based methods described above. If the true relationship between predictor variables and the outcome is linear rather than non-linear, a linear machine learning algorithm like elastic net may fit better than non-linear tree based models. Elastic net is a penalized regression model that performs grouped variable

selection (for details of the elastic net algorithm, see section 5.4.5.3. on page 77). The tuning parameters $\alpha$ and $\lambda$, which control the strength of the elastic net penalty, were selected by 10-fold cross validation. As this machine learning algorithm is restricted to numeric predictors, categorical variables were converted to dummy variables. Again, elastic net requires complete data so missing predictor variables were imputed.

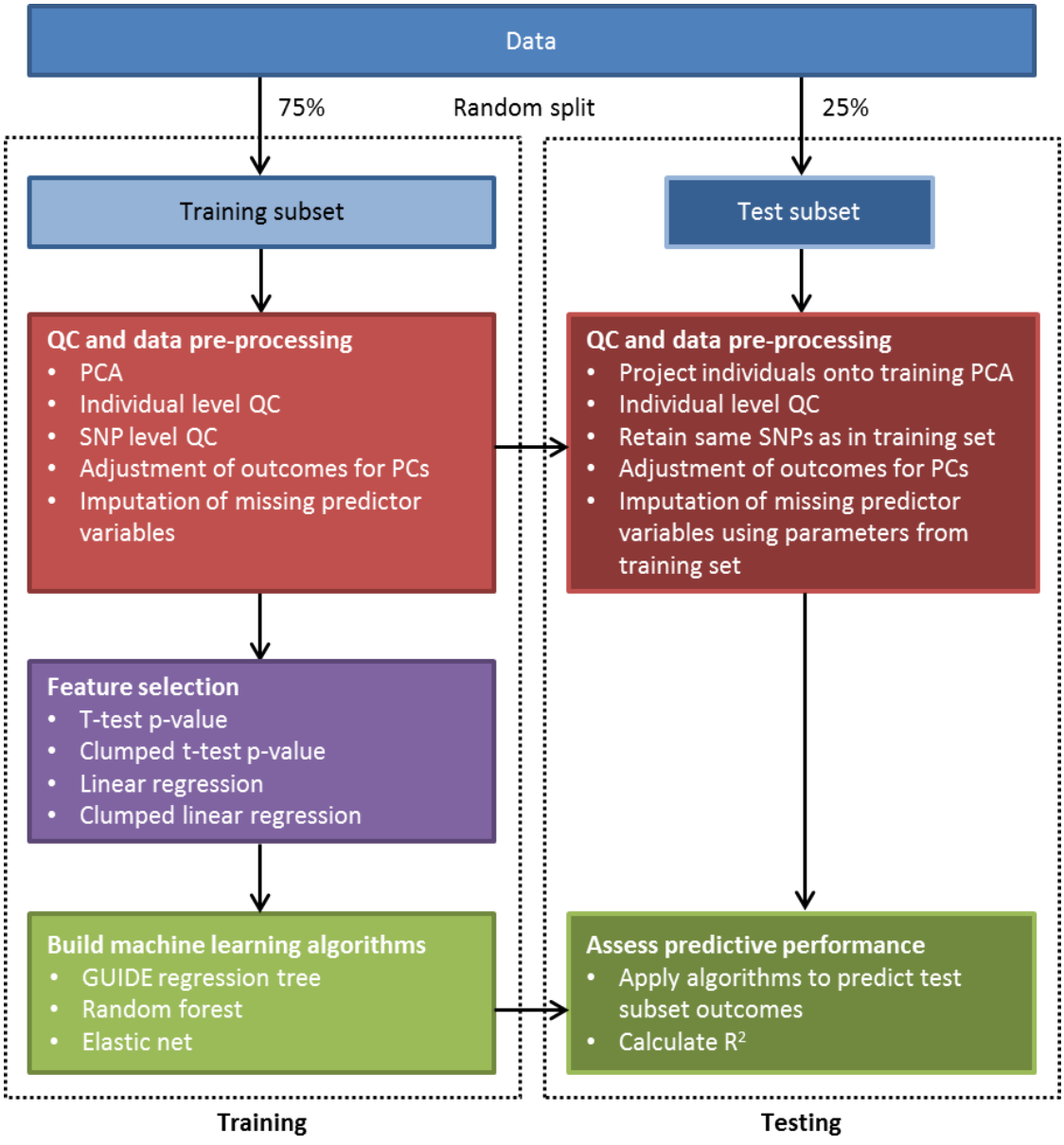



Figure 28. Analysis plan to build machine learning algorithms for the prediction of PGx changes in SBP, HbA1c, FBG, ALT and AST. PCs: principal components; PCA: principal components analysis; QC: Quality Control

*7.2.2.4. Variable selection*

Some machine learning methods, for example elastic net, inherently perform variable selection. Variable selection or feature reduction removes redundant predictors and thus leads to more sparse models and can improve prediction accuracy (Guyon & Elisseeff, 2003). Four different methods were compared to select genetic variables for inclusion in the machine learning algorithms using the training subset. Firstly, the SNPs were tested for association with the outcome variables using t-tests and ranked by increasing *p*-value. The top 10, 50, 100, 250, 500, 1,000, 10,000, 50,000, 100,000, 250,000 and 500,000 SNPs were included in random forest and elastic net models. In GUIDE decision trees, only the top 10, 50, 100, 250, 500 and 1,000 SNPs were used due to software restrictions on the number of predictor variables. Secondly, linear regression models were used to test the association of the SNPs with the outcomes, including clinical predictor variables as covariates. The SNPs were ranked by t-statistic *p*-value and the same numbers of top SNPs were selected. Lastly, both the t-test and regression model ranked SNP lists were clumped using LD ($r^2$ threshold = 0.2) and the top 10, 50, 100, 250, 500, 1,000, 10,000, 50,000, 100,000 and 250,000 independent SNPs were taken forward. The clumped SNP set contained less than 500,000 SNPs, so this bucket was excluded for clumped feature selection. Machine learning algorithms were built using clinical predictors, genetic predictors and clinical and genetic predictors combined. The hyperparameters of the elastic net models were tuned separately for each set of predictor variables.

*7.2.2.5. Sensitivity analysis*

Randomly splitting the data in training and test subsets could by chance lead to extremely precise or poor prediction. Therefore, we performed a sensitivity analysis where the random data splitting, algorithm construction and test set predictive accuracy assessment was repeated 10 times (Fig. 29). This repetition enables evaluation of the reliability of the

predictive performance of the primary analyses. To facilitate automation of the sensitivity analysis, QC, data pre-processing and feature selection were carried out on the entire dataset before the repeated random splitting into training and test sets.

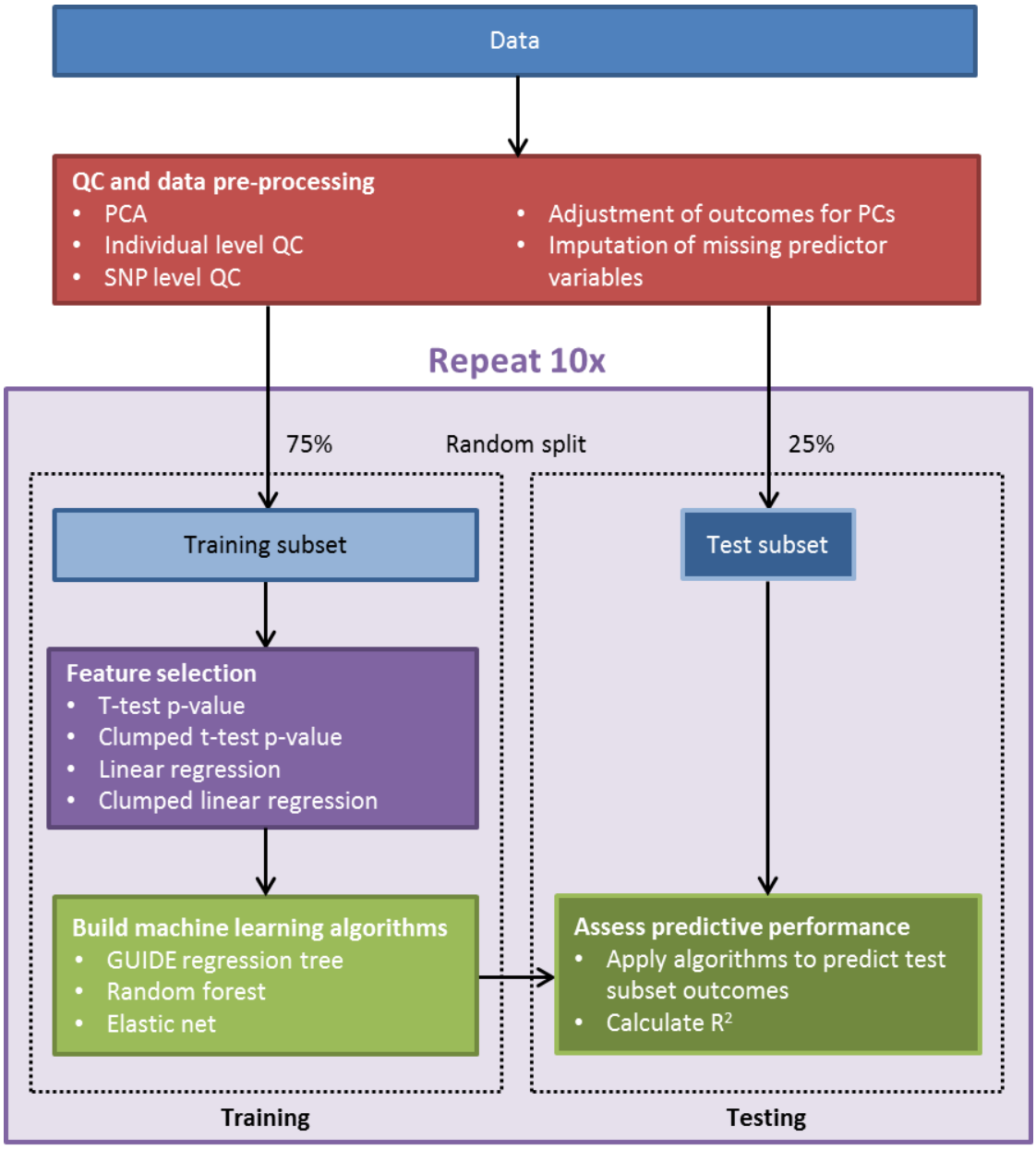



Figure 29. Analysis plan for sensitivity analysis. QC and data pre-processing were performed on the full dataset. Random splitting in training and test subsets, algorithm building and predictive performance assessment was repeated 10 times. PCs: principal components; PCA: principal components analysis; QC: Quality Control

### 7.2.3. Software

GUIDE models were built using GUIDE software, downloadable from the University of Wisconsin-Madison, Department of Statistics website (Loh, 2002, 2009, 2017). Random forest and elastic net analyses were carried out in R version 3.2.5. using the *randomForest* and *glmnet* packages, respectively (Friedman, Hastie, & Tibshirani, 2010; Liaw & Wiener, 2002; R Development Core Team, 2015).

## 7.3. Results

### 7.3.1. Data description

There were 270 participants randomized in the clinical trial, though our study included only the 213 individuals who had at least one observation in the LY2409021 and placebo treatment period and had genome-wide genetic information available. These patients were randomly split into a training subset (n=160) and test subset (n=53). After excluding subjects who failed QC, the number of patients in the training and test samples was 157 and 51, respectively (Table 14). There were slightly more males (56.2%) than females (43.8%) in the sample. The sample was predominantly Caucasian (67.3%) and subjects ranged between 34 and 79 years of age (median = 58). Baseline covariates and mean outcome levels are reported for the training and test sets in Table 15.

Table 14. Number of subjects excluded in the individual-level QC steps in the training and test sets.

| QC step | Training subset (n=160) | Test subset (n=53) |
|---|---|---|
| Individual missingness | 0 | 0 |
| Sex checks | 0 | 0 |
| Cryptic relatedness | 1 | 0 |
| Population stratification | 1 | 2 |
| Heterozygosity and inbreeding | 1 | 0 |
| **Number of subjects retained after QC** | **157** | **51** |

Table 15. Description of outcomes and covariates in training and test data.

| Variable | Training set (n=157) | Test set (n=51) |
|---|---|---|
| **Outcomes** | | |
| Change in SBP in mmHg, mean (s.d.) | 3.08 (8.32) | 0.72 (9.72) |
| Change in HbA1c in %, mean (s.d.) | -0.49 (0.56) | -0.46 (0.5) |
| Change in FBG in mg/dL, mean (s.d.) | -22.61 (32.48) | -23.20 (32.2) |
| Change in ALT in IU/L, mean (s.d.) | 11.27 (17.73) | 12.22 (19.37) |
| Change in AST in IU/L, mean (s.d.) | 7.51 (11.01) | 7.25 (10.23) |
| **Covariates** | | |
| Sex, count (%) | | |
| Male | 87 (55.4) | 30 (58.8) |
| Female | 70 (44.6) | 21 (41.2) |
| Race, count (%) | | |
| Caucasian | 107 (68.2) | 33 (64.7) |
| African American | 20 (12.7) | 3 (5.9) |
| Other | 30 (19.1) | 15 (29.4) |
| Ethnicity, count (%) | | |
| Hispanic | 75 (47.8) | 25 (49.0) |
| Not Hispanic | 71 (45.2) | 24 (47.1) |
| Not reported | 11 (7.0) | 2 (3.9) |
| Age in years, mean (s.d.) | 58.22 (8.91) | 58.57 (7.83) |
| Body weight in kg, mean (s.d.) | 85.83 (14.96) | 89.75 (16.02) |
| Baseline triglycerides in mg/dL, mean (s.d.) | 168.70 (97.19) | 153.08 (66.94) |
| Baseline glucagon in pmol/L, mean (s.d.) | 12.91 (5.79) | 13.3 (7.13) |
| Baseline SBP in mmHg, mean (s.d.) | 129.09 (10.7) | 130.55 (11.02) |
| Baseline HbA1c in %, mean (s.d.) | 7.30 (0.63) | 7.28 (0.58) |
| Baseline FBG in mg/dL, mean (s.d.) | 144.88 (30.82) | 137.23 (27.06) |
| Baseline ALT in IU/L, mean (s.d.) | 28.85 (16.74) | 25.78 (12.2) |
| Baseline AST in IU/L, mean (s.d.) | 24.08 (10.46) | 21.96 (8.35) |

s.d.: standard deviation; IU/L: international units/litre

The participants were genotyped on 745,401 SNPs, of which 41,437 had more than 5% missingness, 78,843 were mono-allelic and 106,231 had MAF less than 2%. Consequently, 518,890 SNPs passed QC and were included as genetic predictor variables in the machine learning algorithms.

The level of missingness in the clinical variables was low: the baseline glucagon and baseline SBP measurements of respectively two and one subject were missing in the training set, and the baseline glucagon measurement of one subject was missing in the test set. Due to this low number of missing values, the KNN and mean imputation methods resulted in virtually identical predictive performance of the clinical random forest and elastic net models (Table 16). We used the KNN imputed clinical variables when combining clinical and genetic predictors.

Table 16. $R^2$ in test set outcome predictions from clinical variable only models using mean and KNN imputation.

| Model | Imputation method | SBP | HbA1c | FBG | ALT | AST |
|---|---|---|---|---|---|---|
| Elastic net | Mean | -0.057 | -0.030 | -0.009 | -0.368 | -0.386 |
| | KNN | -0.057 | -0.030 | -0.008 | -0.369 | -0.386 |
| Random forest | Mean | -0.158 | 0.004 | 0.028 | -0.256 | -0.518 |
| | KNN | -0.172 | 0.008 | 0.018 | -0.262 | -0.527 |

### 7.3.2. Machine learning prediction algorithms

The machine learning algorithms were used to predict the training data outcomes as a measure of model fit. The predictive power of the algorithms was assessed by applying the models to predict the outcomes in the independent test set and calculating the variability explained ($R^2$ statistic). An $R^2$ value of one indicates perfect prediction, whereas $R^2$ equal to zero means a model has no predictive power. Although most models perform well when

predicting the training data, none of the algorithms had meaningful predictive ability in the test data. The method used to select the genetic predictors had a trivial impact on the $R^2$ achieved in all models.

The results of the different machine learning analyses are displayed graphically in Figures 30-35. In each plot, the vertical axis shows the prediction $R^2$, whereas the horizontal axis specifies the number of SNPs that were included in each model. The colours indicate the different feature selection methods that were applied to select SNPs for inclusion, and the plotting symbols represent the type of predictors used in the models (clinical only, genetic only or clinical and genetic combined).The model with clinical predictor variables only (represented by the $*$ symbol) thus corresponds to zero genetic predictors on the left of each graph. The clinical predictors only model and the models including all SNPs are plotted in black as feature selection is not applicable for those models.

*7.3.2.1. GUIDE regression tree*

The prediction accuracy of the constant, simple and stepwise multiple GUIDE regression trees on the training data for each of the five different outcome variables are displayed in Figure 30. In general, GUIDE regression trees including genetic variables (represented by the ● and □ symbols) achieved higher $R^2$ and predicted the training data more accurately than GUIDE models based on clinical predictors only (represented by the $*$ symbol with zero genetic predictors). Furthermore, stepwise multiple regression trees (bottom row of Fig. 30) predicted the training data more closely than the simple and the constant regression trees and resulted in more stable $R^2$ patterns. In the stepwise multiple trees, the predictive performance on the training data improved when more SNPs were included in the models. The constant and simple models showed an inconsistently increasing and decreasing $R^2$ pattern as more genetic predictors were included. For several constant models (top row of Fig. 30) the $R^2$ was equal to zero, which corresponds to no predictive power.

In the independent test data, none of the GUIDE algorithms achieved meaningful predictions (Fig. 31). The maximum $R^2$ reached by any tree is 0.12 though most GUIDE models resulted in negative $R^2$, which indicates the algorithms have no predictive ability. There was large variability in the $R^2$ values, which did not depend on the number of genetic predictors included. We note that in the GUIDE trees predicting FBG, HbA1c and SBP, the algorithms with clinical variables only (represented by the $*$ symbol with zero genetic predictors) predicted better (higher $R^2$) than most models including genetic variables, though not to a relevant degree. Given the instability seen in the results, it is difficult to comment meaningfully on the effect of the different variable selection methods used.

*7.3.2.2. Random forest*

The random forest algorithms predicted the outcomes in the training set closely (Fig. 32). When 10 SNPs and no clinical variables were included as predictors (represented by the ● symbol with 10 genetic predictors), the algorithms achieved lower $R^2$ values, but predictive power was restored when the 10 SNPs were combined with clinical predictors (represented by the □ symbol with 10 genetic predictors). The random forests performed optimally with 100 to 1000 SNPs, and the inclusion of higher numbers of SNPs decreased the $R^2$ slightly. There was no effect of the method used to select the genetic predictors.

In contrast, the random forest algorithms demonstrated little predictive performance in the test data (Fig. 33). The $R^2$ fluctuated around zero and was markedly negative in some AST and SBP algorithms, meaning none of the models predicted the outcomes. Again, there was no difference between the variable selection methods.

*7.3.2.3. Elastic net*

The elastic net algorithms achieved high predictive accuracy in the training data (Fig. 34). The models with clinical predictors only (represented by $*$) performed poorly but the $R^2$

improved steadily when more genetic predictors were included. The inclusion of all 518,890 SNPs in the elastic net models (black ● and □ plotting symbols) reduced the $R^2$ to zero, except for the HbA1c prediction algorithms. All feature selection methods led to similar results.

In the test set, none of the elastic net models had predictive power (Fig. 35). The $R^2$ of the algorithms to predict ALT, AST, HbA1c and SBP improved when more SNPs were added, but did not meaningfully exceed zero. The $R^2$ achieved by the FBG prediction algorithms fluctuated around zero irrespective of the number of genetic predictors. The variable selection methods used to select SNPs had again no notable effect on the results.

*7.3.2.4. Summary of results*

Although the stepwise multiple GUIDE trees, random forest and elastic net algorithms predicted the five different outcomes reliably in the training data (high $R^2$), the models did not retain their predictive power in the test data (zero and negative $R^2$). There was no noticeable difference between the variable selection methods used to select SNPs for inclusion in the models.

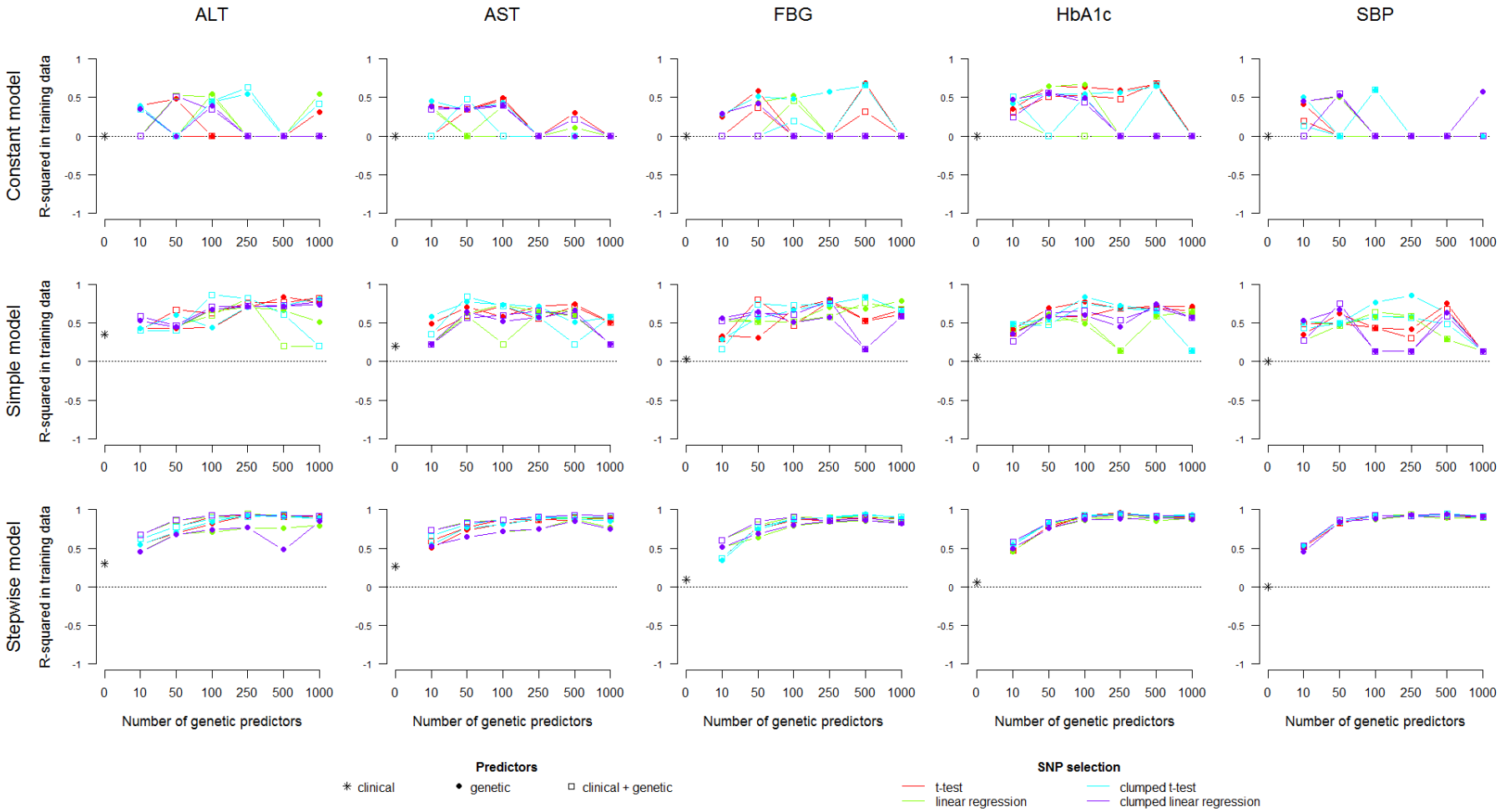



Figure 30. $R^2$ in training set prediction by GUIDE constant, simple and stepwise multiple regression trees (rows) for each of the five outcome variables (columns).

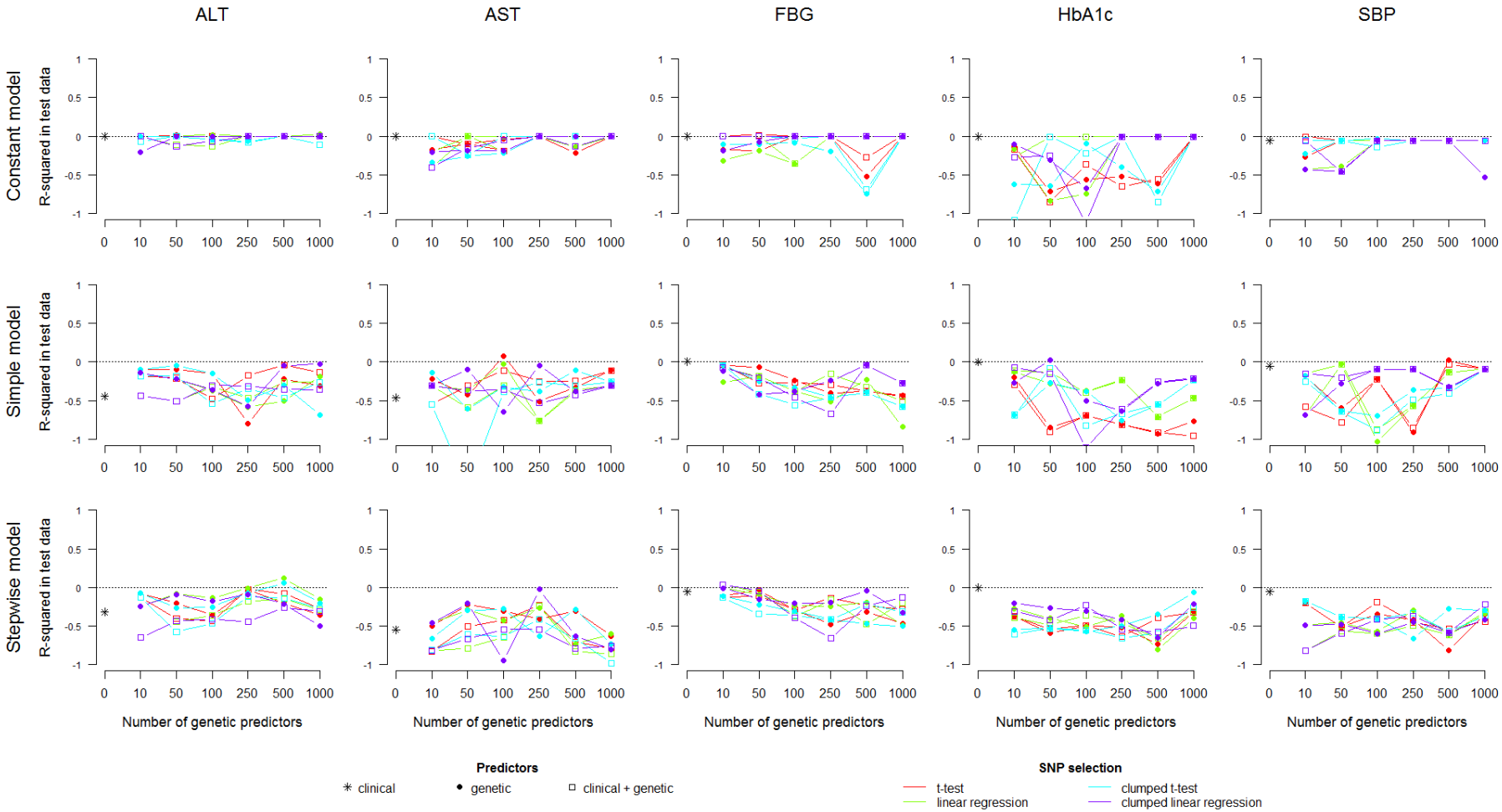



Figure 31. $R^2$ in test set prediction by GUIDE constant, simple and stepwise multiple regression trees (rows) for each of the five outcome variables (columns).

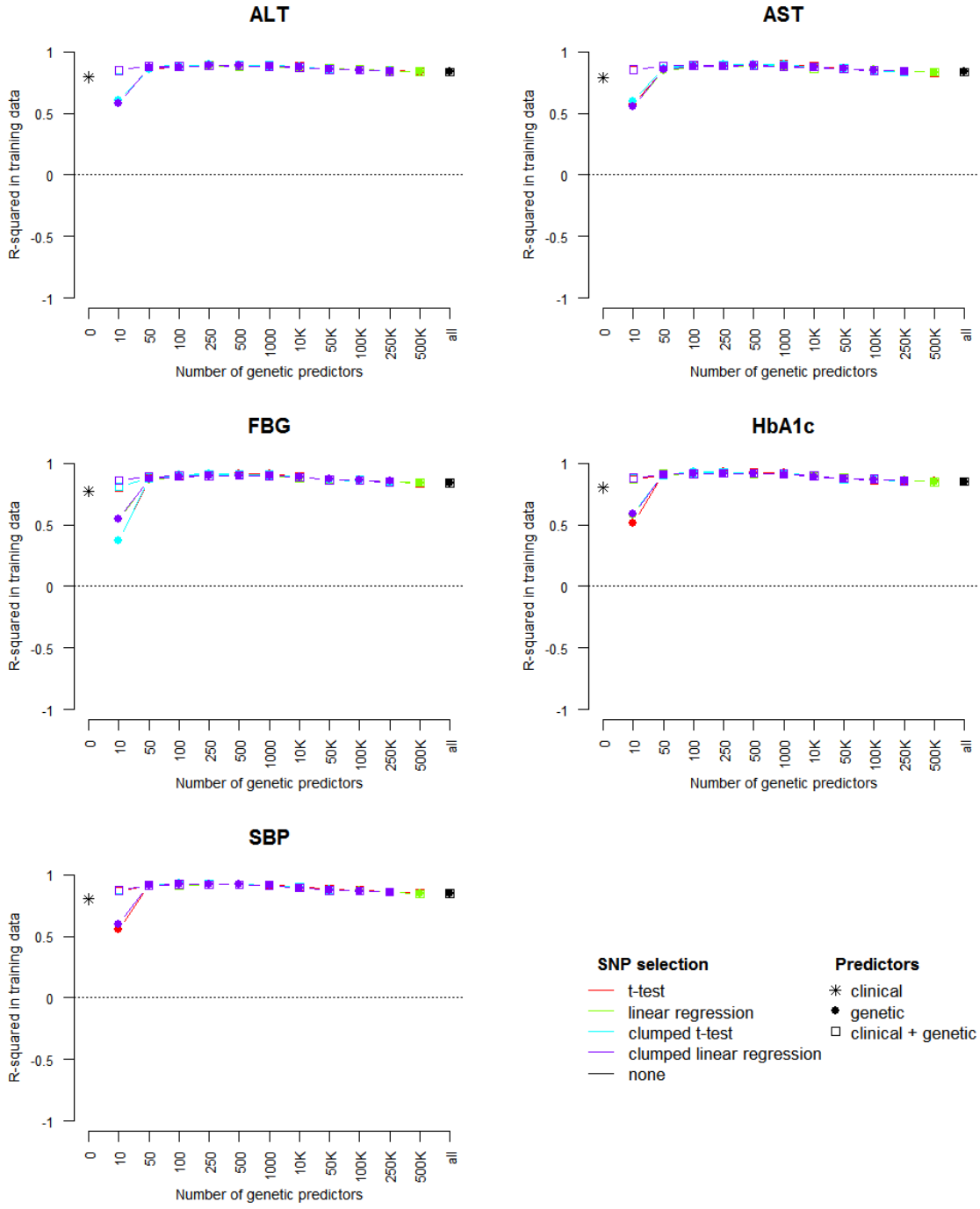



Figure 32. $R^2$ in training set prediction by random forest.

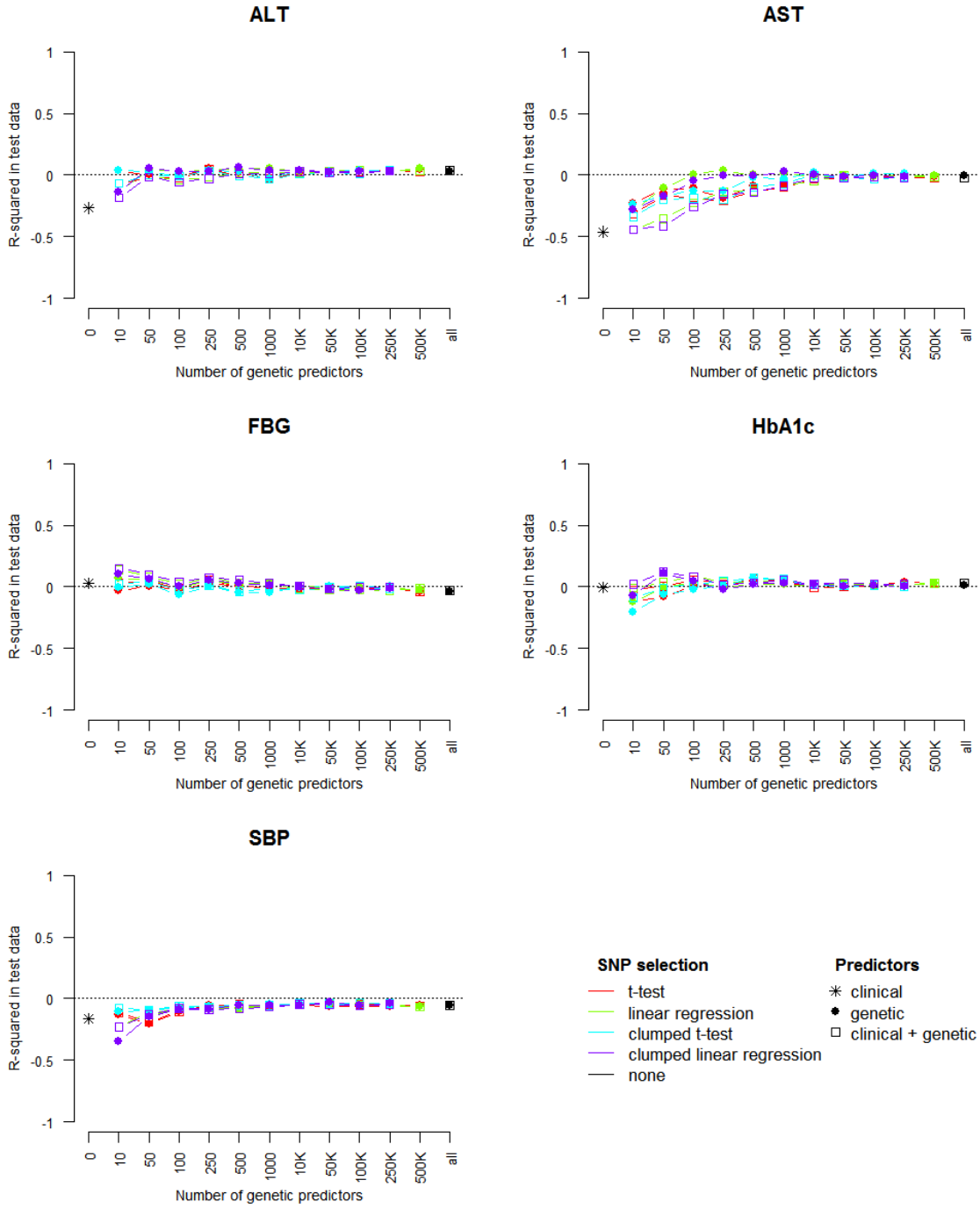



Figure 33. $R^2$ in test set prediction by random forest.

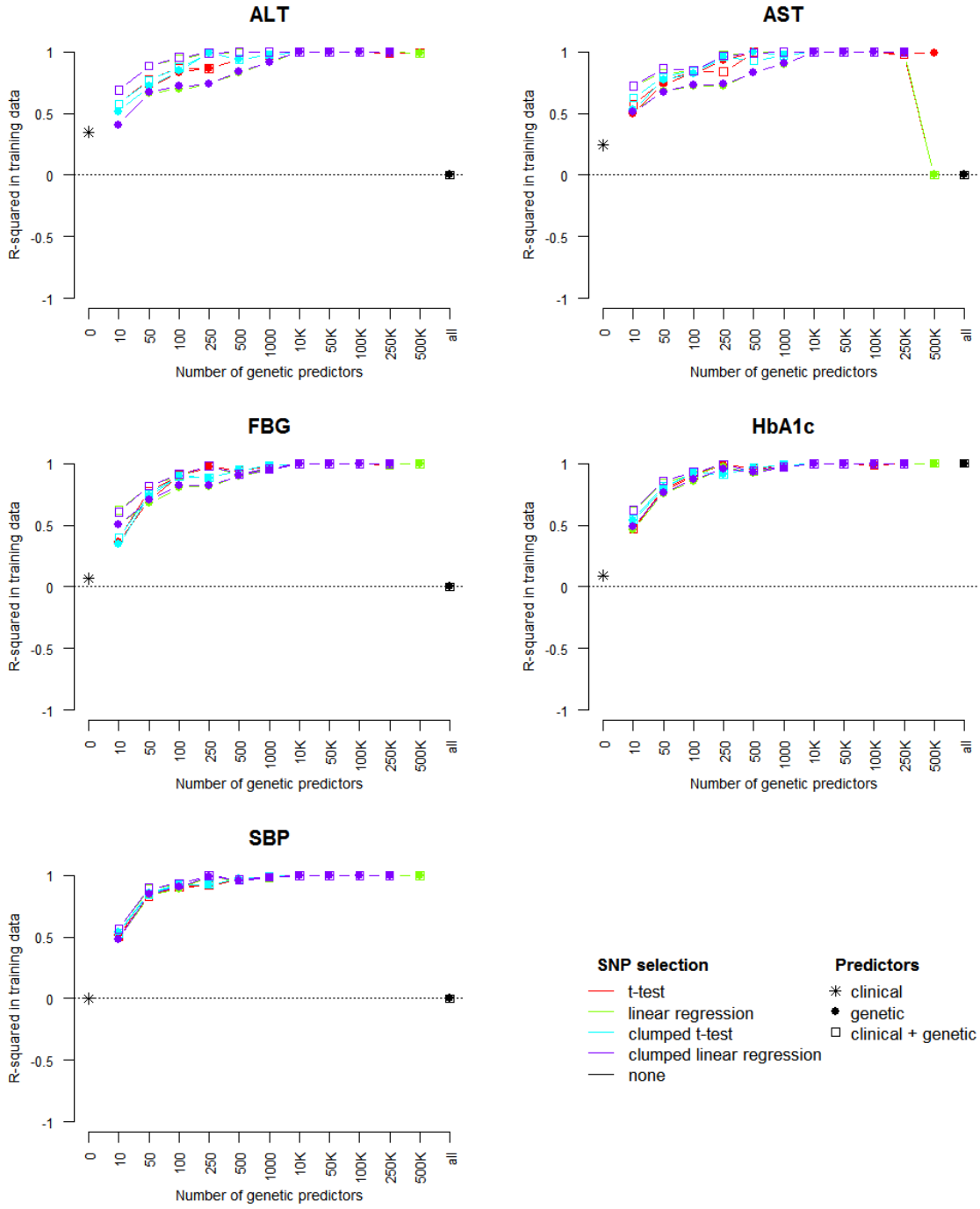



Figure 34. $R^2$ in training set prediction by elastic net.

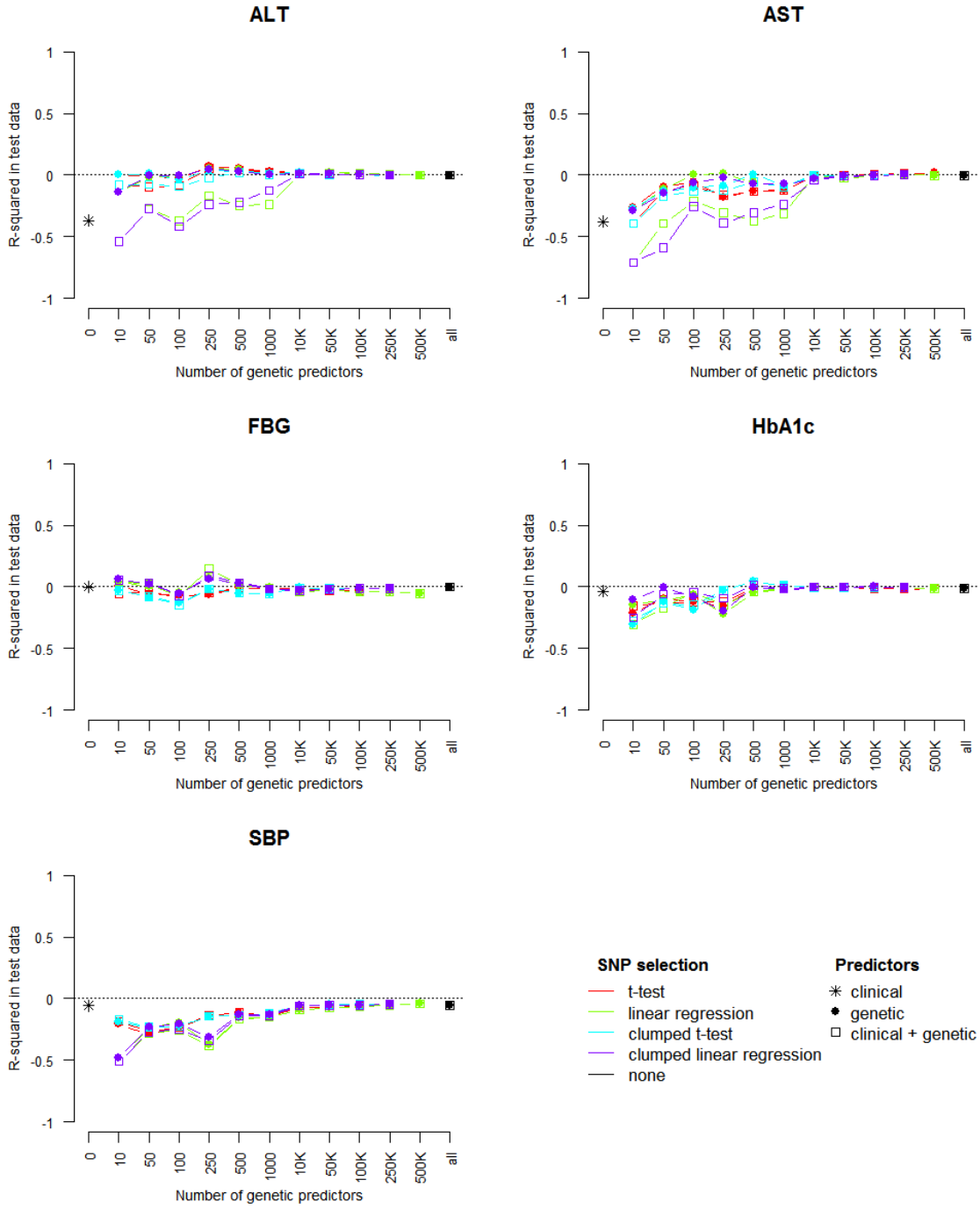



Figure 35. $R^2$ in test set prediction by elastic net.

### 7.3.3. Sensitivity analysis

The sensitivity analysis was carried out to assess the robustness of the results of the primary analyses. Repeating the random data splitting step provides an empirical way to determine the scale of the prediction $R^2$ that can be achieved in this dataset, and circumvents unbalanced data splitting by chance. The $R^2$ values obtained in the primary GUIDE, random forest and elastic net analyses were within the range of those observed in the sensitivity analyses (Fig. 36-41). The mean $R^2$ achieved in the sensitivity analyses followed the same trends as that in the primary analyses, with negative $R^2$ values obtained in the test data set for many of the outcome variables.

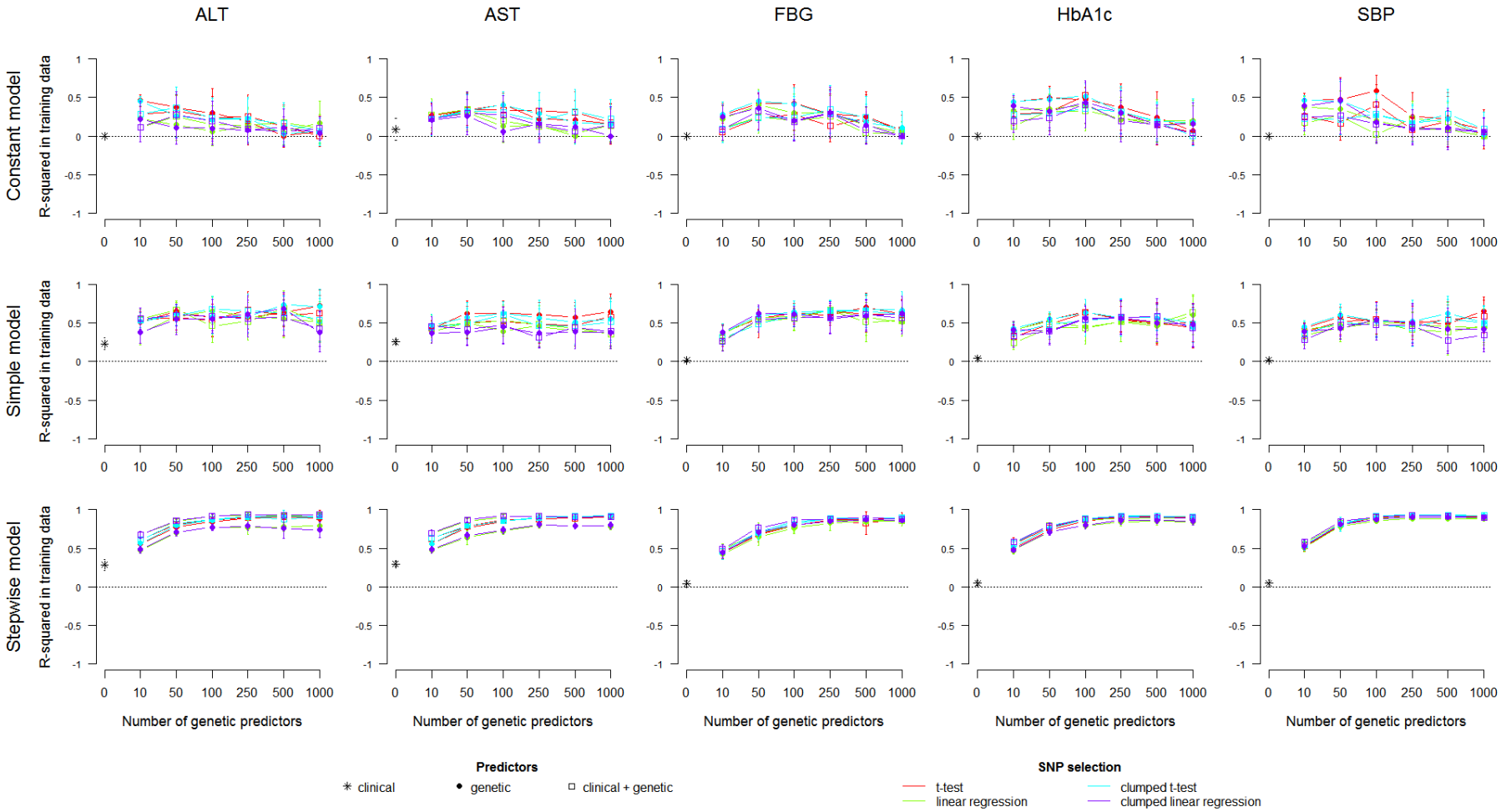



Figure 36. Sensitivity analysis: mean and standard deviation of $R^2$ in training set prediction by GUIDE constant, simple and stepwise multiple regression trees (rows) for each of the five outcome variables (columns).

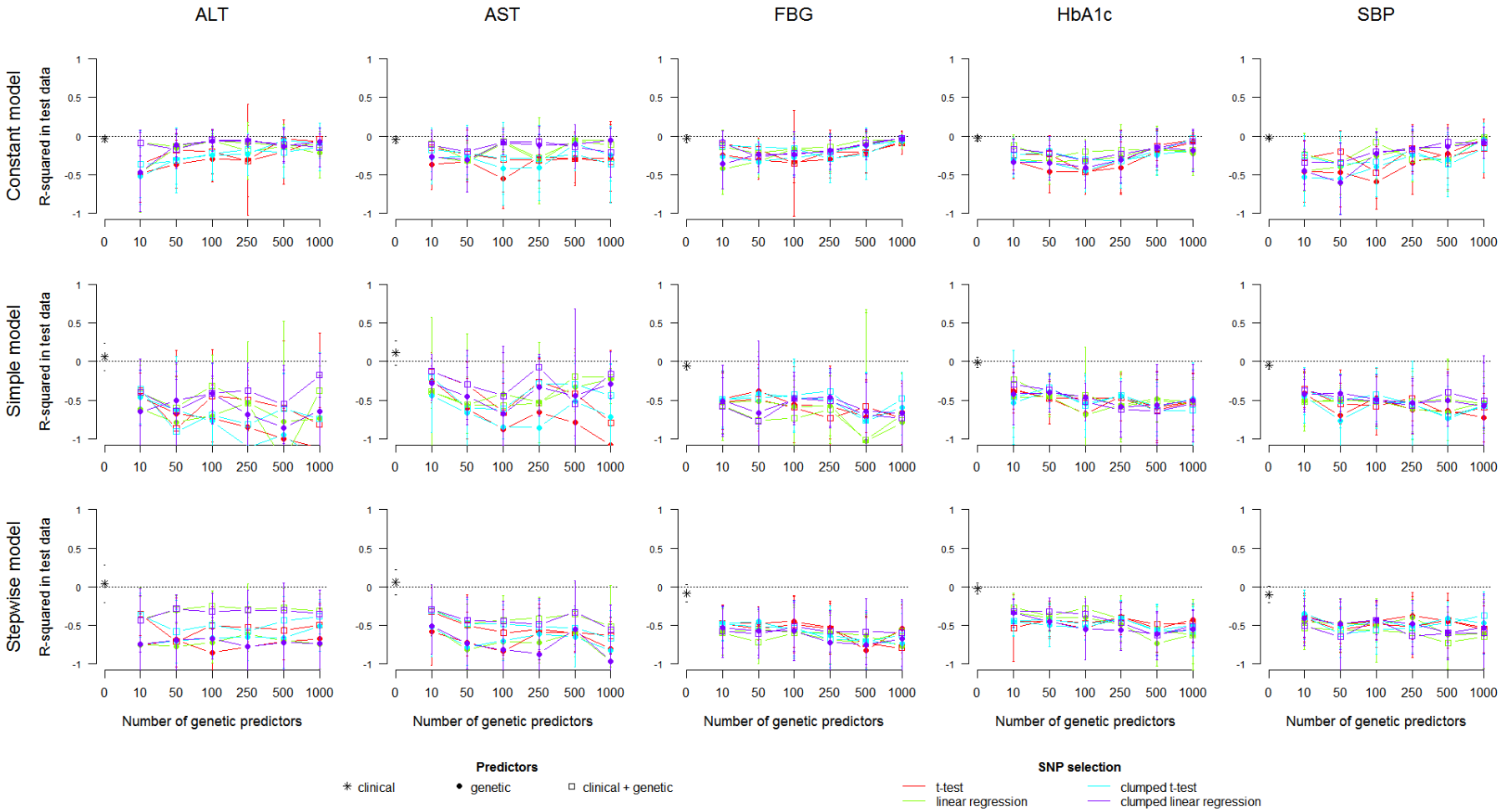



Figure 37. Sensitivity analysis: mean and standard deviation of $R^2$ in test set prediction by GUIDE constant, simple and stepwise multiple regression trees (rows) for each of the five outcome variables (columns).

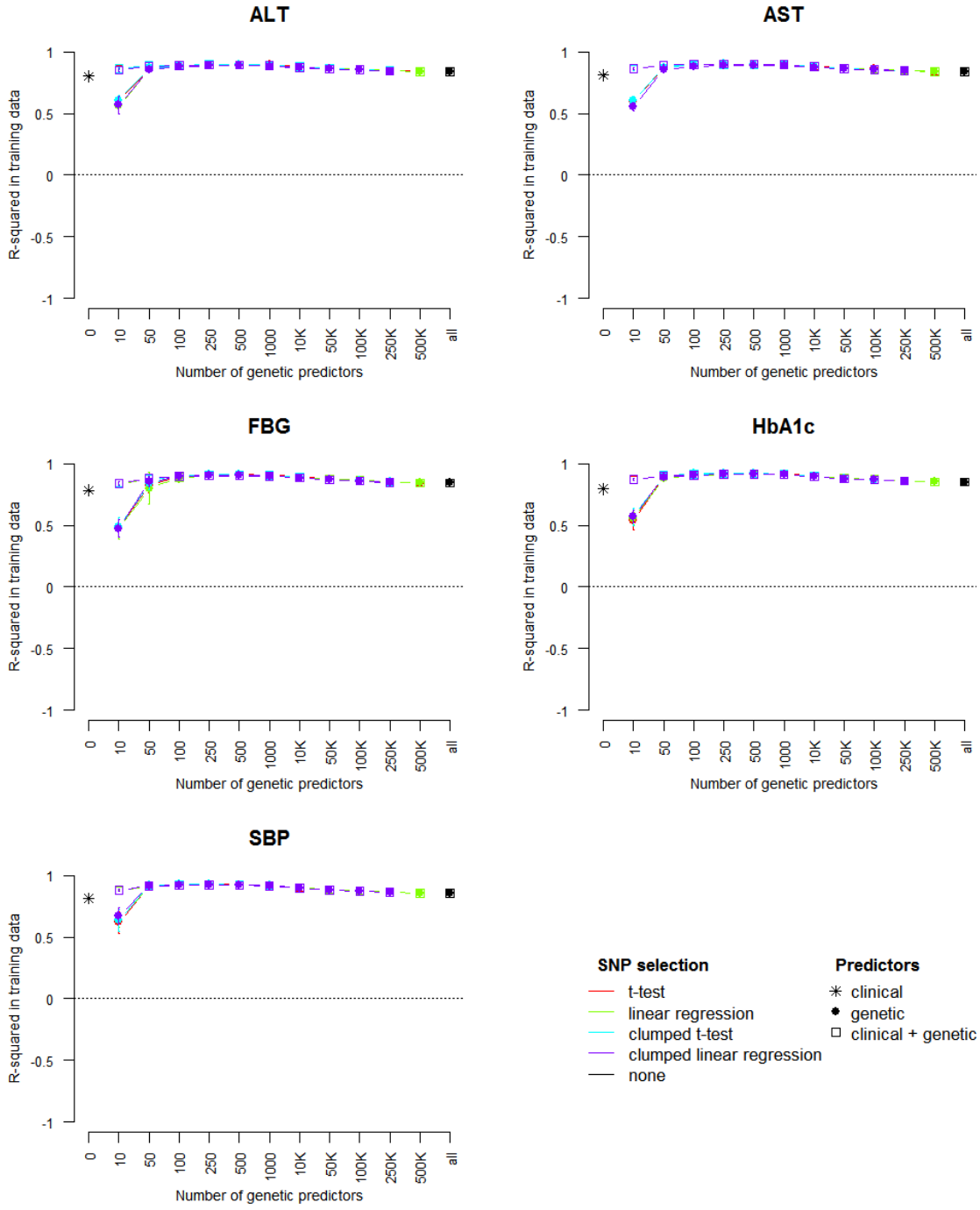



Figure 38. Sensitivity analysis: mean and standard deviation of $R^2$ in trainig set prediction by random forest.

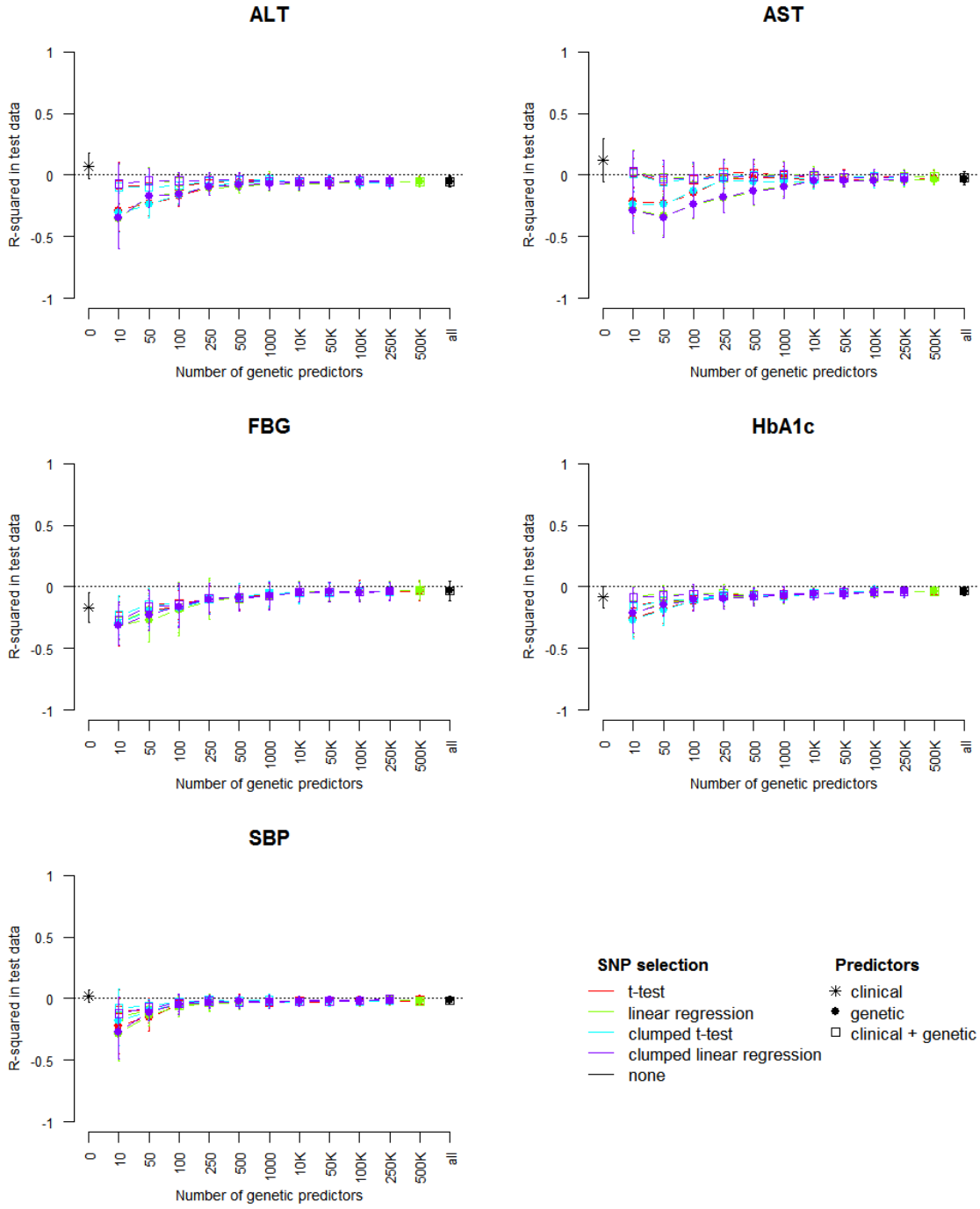



Figure 39. Sensitivity analysis: mean and standard deviation of $R^2$ in test set prediction by random forest.

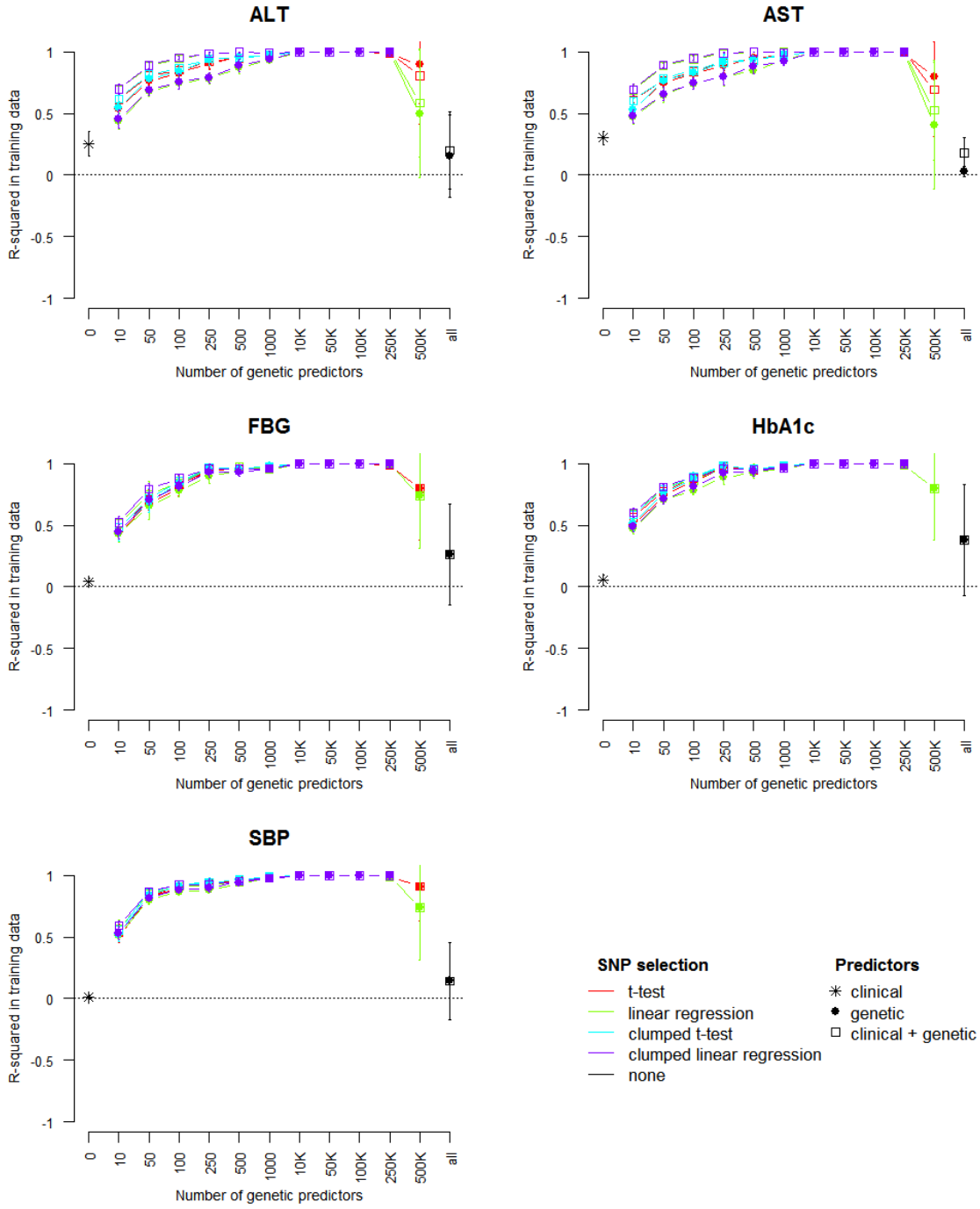



Figure 40. Sensitivity analysis: mean and standard deviation of $R^2$ in training set prediction by elastic net.

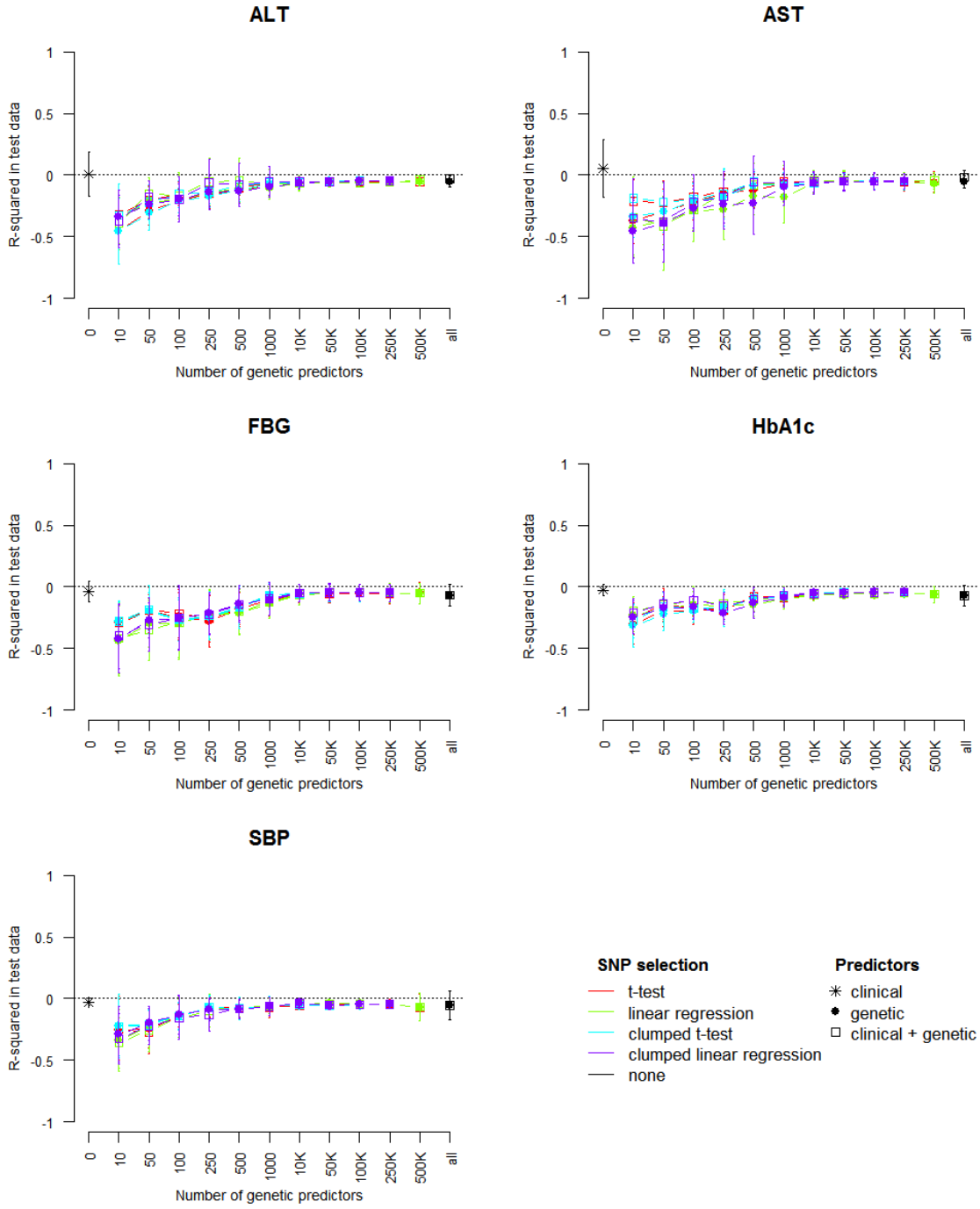



Figure 41. Sensitivity analysis: mean and standard deviation of $R^2$ in test set prediction by elastic net.

## 7.4. Discussion

Machine learning prediction algorithms are a flexible approach for the linear and non-linear analysis of large datasets. These methods have the advantage that they allow the simultaneous analysis of many predictor variables - even more variables than there are data points - unlike traditional statistical models. Therefore, these algorithms are ideal for predictive studies of large genome-wide data, whether or not combined with other variables such as clinical descriptives. Machine learning methods have previously been successfully applied to PGx problems, though often using a small set of pre-selected SNPs as genetic predictor variables (Cosgun et al., 2011; Iniesta et al., 2015; X. Li et al., 2015; Liu et al., 2015).

In this study, the machine learning algorithms applied were not able to predict PGx changes in SBP, HbA1c, FBG, ALT or AST reliably in independent test data. When predictor variables are not associated with the outcome, an $R^2$ statistic around zero is expected. However, many algorithms resulted in a negative prediction $R^2$, indicating that the algorithms do not have any predictive power. The most likely explanation for the negative $R^2$ scores is overfitting of the models to the training data, even though cross-validation was used to select model parameters and prune GUIDE trees. A second factor that may be contributing to poor performance is the sample size (training set $n$=157, test set $n$=51). Typical sample size calculations rely on the assumptions underpinning traditional statistical models and do not apply to machine learning methods. It is thus difficult to determine how large the training dataset needs to be in order to achieve clinically useful predictions if there exist true PGx associations with the outcomes studied. Generally, machine learning methods perform optimally when trained on large samples.

The sensitivity analysis confirmed that predictive GUIDE, random forest or elastic net algorithms could not be built in this dataset. However, it is important to note some

methodological differences between in the primary and sensitivity analyses. QC of the genetic data is difficult to automate, thus for practical purposes QC was performed on the entire dataset, instead of repeating it separately on the training and test data for each iteration of the sensitivity analysis. As some individual-level QC steps use information across subjects (cryptic relatedness and principal components analysis), the consequence is that training and test sets are not strictly independent in the sensitivity analysis. Furthermore, each iteration of data splitting and algorithm building leads to different predictor variables being selected and it is not straightforward to choose the optimally predicting variables. Nevertheless, we do not expect these methodological differences to have a noticeable impact on the prediction accuracy. Our research strategy was to construct prediction algorithms in the primary analysis and to estimate how reliable the test set predictive performance is in the sensitivity analysis.

A few PGx associations with anti-diabetic drugs have been discovered and replicated (Zhou et al., 2016). Metformin, sulfonylureas, thiazolidinediones and gliptins have a PGx association with HbA1c levels, though in our study SNPs could not predict changes in this outcome. The genetic variants described in the PGx literature do not have effects strong enough to guide anti-diabetic treatment (Taber & Dickinson, 2015). Sulphonylureas are the only antidiabetics with a PGx mention on their FDA drug label: patients with glucose 6-phosphate dehydrogenase (*G6PD*) deficiency are at risk of haemolytic anaemia, though genetic testing is not mandatory before initiating sulphonylurea treatment (U.S. Food and Drug Administration, 2013a, 2016b). It is estimated that 415 million people worldwide have diabetes and as this number is expected to increase there is a strong need to develop predictive biomarkers and improve anti-diabetic treatment (International Diabetes Federation, 2015; Maruthur, 2013; World Health Organization, 2016).

The discovery of polygenic PGx biomarkers to guide anti-diabetic treatment in T2DM patients will require large datasets. Machine learning enables the study of multiple genetic

variants simultaneously, on a genome-wide scale and in combination with clinical predictors, and facilitates the development of prediction algorithms to personalize treatments.

# 8. Analysis of pharmacogenetic studies: comparing traditional statistical inference with machine learning

## 8.1. Introduction

Major depressive disorder is a complex disorder with a lifetime prevalence of approximately 15% (Kessler & Bromet, 2013). Patients are treated with a combination of psychotherapy and pharmacotherapy. The response rates for antidepressant is estimated at 55-75%, which means a large group of patients do not respond adequately to treatment (Drago, De Ronchi, & Serretti, 2009).

Although it is estimated that common genetic variants account for 42% of the variance in antidepressant response, few robust PGx associations have been discovered to date (Tansey et al., 2013). Candidate gene studies into PGx associations with antidepressant response were mostly disappointing due to small effect sizes studied in small samples. Some of the candidate genes with PGx associations that have been replicated include antidepressant targets such as the serotonin transporter gene *SLC6A4* and the serotonin receptor gene *HTR2A,* metabolizing enzymes *CYP2D6* and *CYP2C19*, and the drug transporter gene *ABCB1,* though variants in these genes do not achieve significant results in GWAS (Fabbri, Crisafulli, Calabrò, Spina, & Serretti, 2016). Several GWAS of antidepressant response have been conducted, mainly in Caucasian patients, but no SNPs reached genome-wide significance (Biernacka et al., 2015; GENDEP Investigators et al., 2013; Ji et al., 2013). Polygenic risk scores predicted 0.5% of the variance in percentage improvement on a clinician rated depression scale and 1.2% of the variance in remission, though different rating scales and definitions of remission were used in the various samples combined in this study (GENDEP Investigators et al., 2013). Machine learning approaches combining several clinical and genetic predictors explained nearly 4% of the variance in depression severity in

escitalopram and nortriptyline treated patients, which increased to more than 15% of variance explained when both drugs were studied separately (Iniesta et al., 2015). Despite these findings, PGx has not made a relevant impact on depression treatment and more research in this area is necessary.

In this study, we used traditional statistical methods and machine learning algorithms to examine PGx associations with LY2216684 (edivoxetine), a highly selective norepinephrine reuptake inhibitor developed by Eli Lilly as an antidepressant. One clinical trial found that LY2216684 was superior to placebo measured on the Montgomery-Åsberg depression rating scale total score (MADRS-TS) (Pangallo et al., 2011). However, a second trial failed to show that LY2216684 treated patients improved more on the Hamilton depression rating scale than those treated with placebo (Dubé et al., 2010). In 2013, Eli Lilly announced that the drug failed Phase III clinical trials where it was compared to placebo as an add-on to selective serotonin reuptake inhibitor treatment, measured on the MADRS-TS (Eli Lilly and Company, 2013). Although not efficacious in a general population of patients suffering from major depressive disorder, LY2216684 might be beneficial for a subset of patients with certain genetic variants, hence our interest in the PGx analysis of this drug.

## 8.2. Methods

### 8.2.1. Data

The data were collected in a randomized, double-blind clinical trial which found that LY2216684 was superior to placebo for the treatment of major depressive disorder (ClinicalTrials.gov Identifier: NCT00795821) (Pangallo et al., 2011). Participants were assessed at 1, 3, 5, 7 and 10 weeks after the start of the trial. The primary efficacy measure was the clinician rated MADRS-TS, which ranges from 0 to 60 and increases with severity of depression (Montgomery & Asberg, 1979). The MADRS-TS is the sum of 10 questionnaire

items reflecting depression symptoms such as apparent and reported sadness, concentration difficulties and pessimistic thoughts, and each symptom is scored on a 0 to 6 scale. In addition to the outcome measure, the patient characteristics and baseline clinical variables listed in table 17 were included in our study. Patient inclusion criteria and further details of this trial were described previously (Pangallo et al., 2011).

Table 17. Demographic and baseline variables.

| Effect | Description |
|---|---|
| Treatment | LY2216684 or placebo treatment |
| Sex | Sex of the patient |
| Age | Age in years |
| Country | Argentina, Finland, Poland, Russia or United States |
| Alcohol | Whether or not the patient is an alcohol consumer at baseline |
| Tobacco | Whether or not the patient is a tobacco consumer at baseline |
| Baseline MADRS-TS | MADRS-TS at baseline |
| Baseline QIDS | Quick Inventory of Depressive Symptomatology (QIDS) score at baseline |

Although 495 trial participants were randomized, we limited our analysis to 319 white patients for whom genotype data were collected to achieve a more homogeneous sample. 1423 SNPs in 32 candidate genes were genotyped. Standard quality control (QC) procedures were applied to the data set, excluding SNPs with minor allele frequency less than 5% and genotyping rate less than 90%. Hardy-Weinberg Equilibrium tests were carried out and SNPs with significant p-values were flagged but not excluded in QC as all participants suffer from major depressive disorder and deviations from Hardy-Weinberg Equilibrium amongst cases do not necessarily imply genotyping errors. Patients with more than 10% missing SNPs were also excluded from the analyses. The relatively small number of SNPs in our sample did not allow stable principal components analysis or multidimensional scaling to control for population stratification (Price, Zaitlen, Reich, & Patterson, 2010).

### 8.2.2. Statistical methods

Our study tested two hypotheses: (1) that SNPs are associated with change from baseline in MADRS-TS at the end of the trial, and (2) that SNPs are associated with the evolution of change from baseline in MADRS-TS over time.

Both hypotheses were approached using traditional statistical models which investigate each SNP separately, and machine learning techniques that allow the analysis of all SNPs in a single model (Table 18). Although machine learning is often used to develop prediction models, here we applied it to simultaneously analyse a large number of SNPs and identify a set of SNPs associated with changes in the outcome variable. Penalized regression algorithms can be regarded as machine learning analogues of linear regression models and therefore elastic net and linear mixed elastic net were compared to their traditional statistical counterparts, linear regression and linear mixed models, respectively.

Table 18. Analysis approaches used in this study.

| **Modelling approach** | **Outcome variable** | |
|---|---|---|
| | Endpoint change in MADRS-TS | Longitudinal change in MADRS-TS |
| *Traditional statistics*: SNPs modelled separately | Linear regression | Linear mixed model |
| *Machine learning*: SNPs modelled simultaneously | Elastic net | Linear mixed elastic net |

Baseline covariates were included in the traditional and machine learning models through backward selection and the treatment groups were analysed separately as well as combined. The machine learning algorithms used require complete data, thus missing genotypes were imputed using LD in PLINK (Purcell et al., 2007). If LD informed imputation was not possible, genotypes were imputed by the mean allele count of that SNP. There were no missing baseline covariates. Machine learning models were built on the entire

dataset, as the sample size was deemed too small to split the data in separate training and testing subsets.

In addition, we performed a simulation analysis to gain insight into the power that the methods used in the endpoint analyses have to detect genetic associations in this sample.

*8.2.2.1. Endpoint analyses*

To test the first research hypothesis, we investigated the association of SNPs with change in MADRS-TS from baseline at the end of the trial. Linear regression models (see section 5.3.1. on page 61) were used to evaluate the association of each single SNP with the outcome, whereas an elastic net approach (see section 5.4.5.3. on page 77) allowed the analysis of all SNPs simultaneously. For patients who did not complete the trial missing endpoint MADR-TS observations were imputed using a linear mixed model as described below. Baseline covariates to be included in the genetic analyses were determined via backward elimination from a linear regression model including all baseline variables (Table 17) but no genetic variables.

*Linear regression model*

For each SNP a linear regression model was fitted including an intercept, additive genotypic effect and significant baseline covariates. Thus, the model can be represented as

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} + \beta_{SNP} G_i + \varepsilon_i$$

where $y_i$ is the endpoint change in MADRS-TS, $x_1$to $x_m$ are $m$ baseline clinical covariates and $G_i$ is the genotype of a single SNP for the $i$th individual.

Significance of the genetic effect was evaluated using a t-test for each SNP with a Bonferroni type adjustment for multiple testing based on the effective number of independent tests, which accounts for linkage disequilibrium between SNPs (M.-X. Li, Yeung, Cherny, & Sham, 2012).

*Elastic net*

Linear elastic net was used to model the effect of all SNPs simultaneously,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi} + \beta_{SNP1} G_{1i} + \cdots + \beta_{SNPg} G_{gi} + \varepsilon_i$$

where $g$ is the number of SNPs. In addition to SNPs, baseline covariates were included in the elastic net model but their parameter estimates were not penalized. The optimal values for the tuning parameters $\alpha$ and $\lambda$ were selected from a grid by 100 times repeated five-fold cross-validation, using the RMSE

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

as performance measure. We preferred the RMSE over $R^2$ because this is a measure that can easily be carried forward to the longitudinal setting.

*8.2.2.2. Longitudinal analyses*

A study of genetic associations with change in MADRS-TS measurements over time was used to answer the second research hypothesis. SNP main effects and SNP by time (in weeks since start of the trial) interactions were modelled. The latter term captures a potential association with change in outcome over time. Single SNP associations with longitudinal change in MADRS-TS since baseline were analysed using a linear mixed model (see section 5.3.3. on page 63). The joint effect of all genetic variables was investigated with linear mixed elastic net (see section 5.4.5.4. on page 79). Both methods allow unbalanced observations, so no imputation of missing MADRS-TS data was required. Baseline covariates were selected via backward elimination from a linear mixed model including all baseline variables (Table 17) and baseline MADRS-TS and baseline QIDS by time interaction terms. Main effects were always retained if their interaction terms were significant, even if the main effect itself was not significant.

*Linear mixed model*

Linear mixed models are well suited for the analysis of longitudinal data, as they take the correlation between consecutive MADRS-TS observations on the same patient into account. The fixed effects structure of the linear mixed model was built via backward elimination starting from all baseline covariates, baseline MADRS-TS and baseline QIDS by time interaction terms and a quadratic time effect, but no genetic variables. Random intercepts, linear and quadratic time effects were considered as random effects. Thus,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_{\boldsymbol{BL}} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

with $\mathbf{Y}_i$ is the vector of longitudinal changes in MADRS-TS, $\mathbf{X}_i$ is the matrix of baseline fixed effects and $\mathbf{Z}_i$ is the matrix that contains the random intercept and time effects.

Likelihood ratio tests were used to compare the unstructured covariance matrix - which puts no restrictions on the parameters of the random effects covariance matrix - to more parsimonious structures and to test for significance of the random effects (Table 19). The resulting linear mixed model was used to impute missing MADRS-TS observations at week 10 for the endpoint analyses and as a starting point for the longitudinal SNP by SNP analysis.

To test the association of a single SNP with the change in MADRS-TS over time, an additive SNP effect and SNP by time interaction were added (in the matrix $\mathbf{G}_i$) to the fixed effects structure of the mixed model equation,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_{\boldsymbol{BL}} + \mathbf{G}_i\boldsymbol{\beta}_{\boldsymbol{SNP}} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i.$$

The vector $\boldsymbol{\beta}_{\boldsymbol{SNP}}$ has thus 2 components, namely the SNP main effect and the SNP by time interaction. Significance of the SNP by time effect was determined via type III F-tests with a Bonferroni correction for the effective number of SNPs tested (M.-X. Li et al., 2012). REML was used to estimate the parameters of the linear mixed model.

Table 19. Random effects covariance structures.

| Structure | Example |
|---|---|
| Unstructured | $\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$ |
| Independent | $\begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$ |
| Simple | $\begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$ |
| First order auto-regressive | $\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$ |
| Heterogeneous Toeplitz | $\begin{pmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 \\ \rho_1\sigma_1\sigma_2 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 \\ \rho_2\sigma_1\sigma_3 & \rho_1\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$ |

*Linear mixed elastic net*

A linear mixed elastic net model was built starting with the fixed effects structure from the linear mixed model, and adding fixed SNP and SNP by time interaction terms for all genetic variants. The equation for the linear mixed elastic net is the same as for the linear mixed model with the exception that the matrix $\mathbf{G}_i$ and the vector $\boldsymbol{\beta}_{SNP}$ now represent the effects of $g$ SNPs and thus have $2g$ entries.

As was the case in the linear elastic net, baseline covariates were excluded from penalization. Due to computational complexity, some simplifications were made to the random effects structure: the random quadratic time effect was dropped and an independent covariance structure, assuming zero covariance between the random effects, was used for the random effects (Table 19). Nevertheless, we assume that the estimates for

the fixed effects are robust to misspecification of the random effects as is the case in the traditional linear mixed model (Verbeke & Molenberghs, 2000). Model parameters were estimated using ML, as the software used (the *lassop* function of the *MMS* R package) does not support REML (Rohart, 2011; Rohart et al., 2014). Five times repeated five-fold cross-validation was used to select the tuning parameters that minimize the RMSE.

*8.2.2.3. Simulation study*

To gain insight in the power that linear regression and elastic net have to detect genetic associations in this sample we performed a simulation study. We simulated a single SNP associated with the observed change in MADRS-TS at the trial endpoint, appended it to the dataset and tested if the simulated SNP was significantly associated with outcome in the linear regression or selected in the elastic net. SNP effect sizes between 0% and 20% heritability were evaluated, and 1000 simulations were performed at each effect size. The power to detect genetic effects of a given effect size was defined as the proportion of simulations where the simulated SNP was significant (linear regression) or selected (elastic net). Although simulation is a standardly used approach to determine the power of a novel analysis method in genetic studies, most studies simulate phenotypic values from genotypes. Here, we wished to use existing phenotype values (change in MADRS-TS from baseline), and simulate genotype. No standard methods or software exists for this, and full algebraic details of the simulation framework are given in the following paragraphs. A comparable simulation analysis for the longitudinal analyses was not performed due to the computational burden of the linear mixed elastic net.

*Algebraic details of simulation study*

To simulate a SNP that is associated with the change in MADRS-TS at the endpoint, the following linear model was assumed:

$$y_i = \alpha + \beta G_i + \varepsilon_i \quad with \;\; \varepsilon_i \sim N(0, \sigma_e^2) \tag{Eq. 1}$$

where $y_i$ is the outcome at the end of the trial, $\alpha$ is the intercept, $G_i$ is the genotype and $\beta$ its effect size. We assume an additive genetic effect, so $G_i$ is coded as 0, 1 or 2 minor allele counts and the minor and major allele frequencies are $p$ and $q$, respectively.

The total variance can be decomposed into the variance explained by the genotype and the residual error variance,

$$\sigma^2 = \sigma_G^2 + \sigma_e^2 \tag{Eq. 2}$$

The variance explained by the genotype can be computed from the additive effect size and the minor allele frequency:

$$\sigma_G^2 = 2\beta^2 pq \tag{Eq. 3}$$

Bayes' theorem allows us to calculate the probability for each genotype conditional on the observed change in the outcome,

$$P(G = g|y) = \frac{P(y|G = g)P(G = g)}{P(y)} \tag{Eq. 4}$$

Thus, to calculate $P(G = g|y)$ the following quantities are required:

**a) Probability of the genotype G:** $P(G = g)$

Assuming Hardy-Weinberg equilibrium,

$$\begin{aligned} P(G = 0) &= q^2 \\ P(G = 1) &= 2pq \\ P(G = 2) &= p^2 \end{aligned} \tag{Eq. 5}$$

**b) Probability of observing the outcome y conditional on the genotype G:** $P(y|G=g)$

$y|G=g$ follows a normal distribution with mean:

$$\begin{aligned} E(y|G=0) &= E(\alpha+\varepsilon) = \alpha \\ E(y|G=1) &= E(\alpha+\beta+\varepsilon) = \alpha+\beta \\ E(y|G=2) &= E(\alpha+2\beta+\varepsilon) = \alpha+2\beta \end{aligned} \quad \text{(Eq. 6)}$$

From the law of total probability, we infer the value of $\alpha$,

$$\begin{aligned} &E(y) = \mu \\ &= E(y|G=0)P(G=0) + E(y|G=1)P(G=1) + E(y|G=2)P(G=2) \\ &= \alpha q^2 + (\alpha+\beta)2pq + (\alpha+2\beta)p^2 \\ &= \alpha(q^2+2pq+p^2) + \beta(2pq+2p^2) \\ &= \alpha + 2\beta p(q+p) \\ &= \alpha + 2\beta p \end{aligned} \quad \text{(Eq. 7)}$$

Thus, $\alpha = \mu - 2\beta p$ and the means of the conditional distributions of $y$ are

$$\begin{aligned} &E(y|G=0) = E(\alpha+\varepsilon) = \alpha = \mu - 2\beta p \\ &E(y|G=1) = E(\alpha+\beta+\varepsilon) = \alpha+\beta = \mu - \beta(2p-1) = \mu - \beta(p-q) \\ &E(y|G=2) = E(\alpha+2\beta+\varepsilon) = \alpha+2\beta = \mu - 2\beta(p-1) = \mu + 2\beta q \end{aligned} \quad \text{(Eq. 8)}$$

Assuming that the variance of $y$ conditional on $G$ is constant, it is equal to $\sigma_e^2$.

Combining equations 2 and 3 gives

$$\sigma_e^2 = \sigma^2 - 2\beta^2 pq \quad \text{(Eq. 9)}$$

In equation 8 and 9, the quantities $\mu$ and $\sigma^2$ can be approximated by the mean $\overline{y}$ and variance $\hat{\sigma}^2$ of the observed outcomes.

Thus, the conditional distributions of $y$ are the following,

$$\begin{aligned} y|G=0 &\sim N(\mu - 2\beta,\ \sigma^2 - 2\beta^2 pq) \\ y|G=1 &\sim N(\mu\ - \beta(p-q),\ \sigma^2 - 2\beta^2 pq) \\ y|G=2 &\sim N(\mu\ + 2\beta q,\ \sigma^2 - 2\beta^2 pq) \end{aligned} \quad \text{(Eq. 10)}$$

and $P(y|G=g)$ is simply the probability density function of these normal distributions,

$$P(y|G=0) = \frac{1}{\sqrt{2\pi(\sigma^2 - 2\beta^2 pq)}} e^{-\frac{(y-\mu+2\beta)}{2(\sigma^2-2\beta^2 pq)}}$$

$$P(y|G=1) = \frac{1}{\sqrt{2\pi(\sigma^2 - 2\beta^2 pq)}} e^{-\frac{(y-\mu+\beta(p-q))}{2(\sigma^2-2\beta^2 pq)}} \quad \text{(Eq. 11)}$$

$$P(y|G=2) = \frac{1}{\sqrt{2\pi(\sigma^2 - 2\beta^2 pq)}} e^{-\frac{(y-\mu-2\beta q)}{2(\sigma^2-2\beta^2 pq)}}$$

**c) Probability of the outcome $y$:** $P(y)$

The law of total probability allows straightforward calculation of $P(y)$,

$$P(y) = P(y|G=0)P(G=0) + P(y|G=1)P(G=1) + P(y|G=2)P(G=2) \quad \text{(Eq. 12)}$$

the elements of which have all been defined previously (Eq. 5 and 11).

Thus, by specifying the genetic effect size ($\beta$) and minor allele frequency ($p$), $P(G=g|y)$ can be calculated for each value *g* (Eq. 4). For each subject in the dataset, the probabilities of the genotypes conditional on the observed outcome were calculated. The SNP genotype was then simulated from a multinomial distribution.

### 8.2.3. Software

Linear regression analyses were performed using PLINK, version 1.07 (Purcell et al., 2007). Linear mixed model analyses were carried out using proc mixed in SAS software, version 9.3 (SAS Institute, Cary, NC). Linear elastic net and linear mixed elastic net algorithms were modelled in R version 3.1.2 using the *caret* and *glmnet*, and *MMS* packages, respectively (Friedman et al., 2010; Kuhn, 2008; R Development Core Team, 2015; Rohart, 2016; Rohart et al., 2014). The simulation study was also carried out in R.

## 8.3. Results

### 8.3.1. Data description

No SNPs had less than 90% genotyping rate, but 285 SNPs were removed because their minor allele frequencies were less than 5%, reducing the number of SNPs analysed to 1,138 (Table 20). Five SNPs deviated significantly from Hardy-Weinberg Equilibrium (rs908867, rs4713908, rs2601608, rs512131 and rs11186300) but where nevertheless included in the analyses. No individuals had more than 90% SNPs missing, thus all 319 patients were included in the analyses. There were no significant differences in demographics or baseline characteristics between the LY2216684 and placebo treated patients (Table 21).

### 8.3.2. Endpoint analyses

*8.3.2.1. Outcome variable*

For 10% of patients the week 10 outcome observation was missing and imputed using a linear mixed model. The mean and median change in MADRS-TS at the trial endpoint were -11.20 and -11, respectively (Fig.42). It should be noted that negative outcome values reflect improvements in the patient's condition as lower MARDS-TS scores indicate less severe depression symptoms. At the end of the trial period, 35.4% of patients achieved response to treatment (decrease in MADRS-TS≥50%) and 21.0% reached remission (MADRS-TS≤10).

*8.3.2.2. Baseline covariates*

The linear regression model including only baseline covariates revealed that baseline MADRS-TS, patient age and treatment were significantly associated with change in MADRS-TS at week 10 in the full sample (Table 22). In the treatment specific analyses, only baseline MADRS-TS remained significantly associated with the outcome variable. These variables were thus included in the linear regression and elastic net analyses. As negative outcome

scores reflect patient improvement, negative parameter coefficients imply a beneficial effect, whereas positive coefficients denote a disadvantageous effect on patient wellbeing.

Table 20. List of candidate genes studied and the number of SNPs typed in each gene.

| Gene | Chromosome | Number of SNPs before QC | Number of SNPs after QC |
|---|---|---|---|
| *KCNK2* | 1 | 238 | 192 |
| *RGL1* | 1 | 2 | 2 |
| *ADRA2B* | 2 | 6 | 5 |
| *DRD3* | 3 | 48 | 32 |
| *ADRA2C* | 4 | 2 | 2 |
| *ADRB2* | 5 | 15 | 13 |
| *DRD1* | 5 | 11 | 11 |
| *HTR1A* | 5 | 6 | 4 |
| *NR3C1* | 5 | 70 | 61 |
| *FKBP5* | 6 | 61 | 54 |
| *UST* | 6 | 5 | 5 |
| *DDC* | 7 | 116 | 95 |
| *LEP* | 7 | 1 | 1 |
| *ADRA1A* | 8 | 155 | 123 |
| *DBH* | 9 | 41 | 35 |
| *ADRA2A* | 10 | 10 | 8 |
| *ADRB1* | 10 | 6 | 6 |
| *HTR7* | 10 | 65 | 62 |
| *BDNF* | 11 | 37 | 27 |
| *DRD2* | 11 | 69 | 65 |
| *DRD4* | 11 | 3 | 3 |
| *GRIK4* | 11 | 6 | 6 |
| *HTR3A* | 11 | 26 | 14 |
| *TH* | 11 | 8 | 8 |
| *TPH1* | 11 | 26 | 25 |
| *HTR2A* | 13 | 128 | 114 |
| *SLC6A2* | 16 | 117 | 79 |
| *SLC6A4* | 17 | 36 | 24 |
| *GRIK1* | 21 | 1 | 1 |
| *COMT* | 22 | 36 | 25 |
| *MAOA* | X | 34 | 28 |
| *MAOB* | X | 38 | 8 |
| **Total count** | **-** | **1423** | **1138** |

Table 21. Demographics and baseline clinical characteristics by treatment group.

| **Characteristics** | **LY2216684 (n=157)** | **Placebo (n=162)** | ***p*-value** |
|---|---|---|---|
| Age in years, mean (s.d.) | 45.33 (11.89) | 46.93 (11.53) | 0.22 |
| Sex, count (%) | | | 0.41 |
| Male | 55 (35.03) | 64 (39.51) | |
| Female | 102 (64.97) | 98 (60.49) | |
| Country, count (%) | | | 0.94 |
| Argentina | 10 (6.37) | 11 (6.79) | |
| Finland | 36 (22.93) | 39 (24.07) | |
| Poland | 36 (22.93) | 33 (20.37) | |
| Russia | 8 (5.1) | 6 (3.7) | |
| United States | 67 (42.68) | 73 (45.06) | |
| Alcohol use, count (%) | | | 0.32 |
| No | 109 (69.43) | 104 (64.20) | |
| Yes | 48 (30.57) | 58 (35.80) | |
| Tobacco use, count (%) | | | 0.13 |
| No | 114 (72.61) | 105 (64.81) | |
| Yes | 43 (27.39) | 57 (35.19) | |
| Baseline MADRS-TS, mean (s.d.) | 29.31 (3.97) | 29.95 (4.12) | 0.16 |
| Baseline QIDS, mean (s.d.) | 14.34 (3.82) | 15.05 (3.55) | 0.09 |

s.d.: standard deviation

Table 22. Significant covariates in baseline regression model.

| **Parameter** | **Full sample (n=319)** | | **LY2216684 (n=157)** | | **Placebo (n=162)** | |
|---|---|---|---|---|---|---|
| | Parameter estimate | *p*-value | Parameter estimate | *p*-value | Parameter estimate | *p*-value |
| LY2216684 treatment | -3.84 | <.0001 | - | - | - | - |
| Age | 0.08 | 0.049 | - | - | - | - |
| Baseline MADRS-TS | -0.69 | <.0001 | -0.64 | <.0001 | -0.68 | <.0001 |

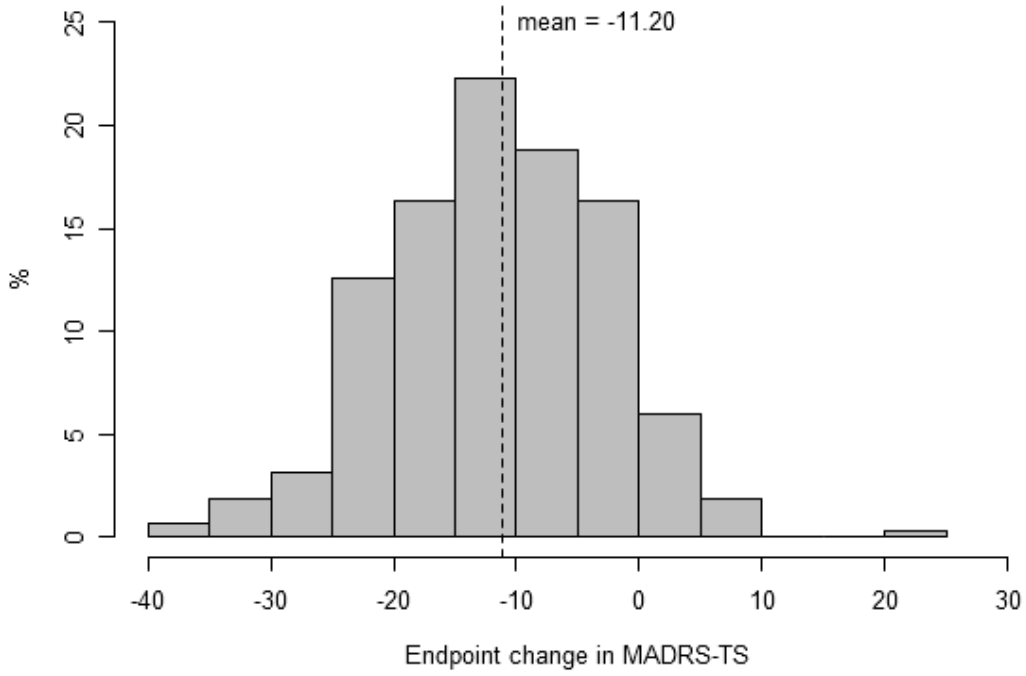



Figure 42. Histogram of change in MADRS-TS at the end of the trial.

*8.3.2.3. Linear regression analysis*

No SNPs were significantly associated with change in MADRS-TS at the end of the trial in either dataset after correction for multiple testing (Table 23). The effective number of tests were 383 in the full sample, 371 in the LY2216684 treated subset and 364 in the placebo treated subset, so significance thresholds were $1.31*10^{-4}$, $1.35*10^{-4}$ and $1.37*10^{-4}$ respectively.

SNPs in *ADRA1A* and *HTR2A* achieve low *p*-values across different analyses (Table 20). For each sample, SNPs within the same gene represent a single association, but there is little overlap between the different analyses. The full sample SNPs in *HTR2A* are in strong LD ($r_{LD}^2$>0.82), as are the SNPs in *ADRA1A* ($r_{LD}^2$>0.94). The *ADRA1A* SNPs from the LY2216684 analysis are in LD with each other ($r_{LD}^2$>0.99), but not with the full sample SNPs in the same gene ($r_{LD}^2$<0.15). Similarly, the *HTR2A* SNPs from the placebo analysis are in strong LD with each other ($r_{LD}^2$>0.92), but in much weaker LD with the full sample *HTR2A* SNPs ($r_{LD}^2$<0.58).

In both genes a large number of SNPs was genotyped relative to the other genes in the analysis.

*8.3.2.4. Elastic net analysis*

The full sample elastic net with baseline covariates only had a mean cross-validation RMSE of 8.30. When SNPs were added, the full sample elastic net selected 23 SNPs for inclusion in the model, reducing the mean cross-validation RMSE to 8.27 (Table 24). The treatment specific elastic net models did not retain any SNPs. Looking at the tuning parameters of the full sample elastic net, we note that the model chosen by five-fold cross-validation was closer to a ridge regression than a lasso ($\alpha$=0.2) and that the penalty was strong to encourage variable selection ($\lambda$=4.8).

The SNPs identified in the full sample analysis showed consistency with the linear regression results (Table 25). From the 30 SNPs with lowest *p*-values in linear regression, 20 were selected by the elastic net, including the top 8 SNPs. All SNPs selected by the elastic net ranked 50$^{th}$ or higher in the linear regression results sorted by ascending *p*-value.

Inspection of the linkage disequilibrium between the SNPs revealed that the elastic net indeed performed grouped variable selection in some groups of correlated SNPs, but not in all. For example, on chromosome 11, SNPs rs7128320, rs12800734 and rs2276319 are highly correlated ($r_{LD}^2$>0.99) and all 3 were included in the elastic net. On the contrary, on chromosome 10, rs2420203 is in strong LD ($r_{LD}^2$>0.96) with 2 other SNPs but it is the only one of that group that was selected.

Table 23. SNPs with the five lowest *p*-values in the full sample (main effect and SNP by treatment interaction models), LY2216684 and placebo linear regression analyses. For the LY2216684 and placebo analyses we also report the top 5 SNPs from the full sample SNP by treatment interaction model.

| SNP | Reference allele | Gene | Chromosome | Position | Parameter estimate (s.e.) | *p*-value |
|---|---|---|---|---|---|---|
| Full sample analysis (n=319) - SNP main effect | | | | | | |
| rs17288723 | G | *HTR2A* | 13 | 46883558 | 3.01 (0.89) | 0.0008 |
| rs731779 | C | *HTR2A* | 13 | 46877903 | 2.71 (0.83) | 0.0013 |
| rs505138 | A | *ADRA1A* | 8 | 26869465 | 3.66 (1.13) | 0.0013 |
| rs17287961 | A | *HTR2A* | 13 | 46862976 | 2.86 (0.89) | 0.0015 |
| rs580644 | A | *ADRA1A* | 8 | 26862973 | 3.63 (1.14) | 0.0015 |
| Full sample analysis (n=319) – SNP by treatment interaction | | | | | | |
| rs11841433 | A | *HTR2A* | 13 | 47434733 | -9.06 (2.76) | 0.0011 |
| rs1328674 | A | *HTR2A* | 13 | 47441707 | -8.89 (2.78) | 0.0015 |
| rs7326071 | A | *HTR2A* | 13 | 47438668 | -8.93 (2.80) | 0.0016 |
| rs3785151 | C | *SLC6A2* | 16 | 55712519 | 5.49 (1.87) | 0.0036 |
| rs2025296 | A | *HTR2A* | 13 | 47463819 | -6.26 (2.17) | 0.0042 |
| LY2216684 analysis (n=157) - SNP main effect | | | | | | |
| rs12121815 | G | *KCNK2* | 1 | 215187382 | 5.22 (1.78) | 0.0038 |
| rs511662 | G | *ADRA1A* | 8 | 26848689 | -2.72 (1.00) | 0.0071 |
| rs12521436 | A | *NR3C1* | 5 | 143438042 | -3.18 (1.19) | 0.0080 |
| rs577366 | G | *ADRA1A* | 8 | 26839848 | -2.57 (1.00) | 0.0111 |
| rs13282250 | C | *ADRA1A* | 8 | 26836875 | -2.57 (1.00) | 0.0111 |
| rs11841433 | A | *HTR2A* | 13 | 47434733 | -4.06 (1.65) | 0.0149 |
| rs1328674 | A | *HTR2A* | 13 | 47441707 | -3.94 (1.68) | 0.0203 |
| rs7326071 | A | *HTR2A* | 13 | 47438668 | -4.02 (1.71) | 0.0200 |
| rs3785151 | C | *SLC6A2* | 16 | 55712519 | 1.37 (1.31) | 0.2961 |
| rs2025296 | A | *HTR2A* | 13 | 47463819 | -3.26 (1.43) | 0.0242 |
| Placebo analysis (n=162) - SNP main effect | | | | | | |
| rs2770298 | G | *HTR2A* | 13 | 46872712 | 3.50 (1.03) | 0.0009 |
| rs2770296 | G | *HTR2A* | 13 | 46866425 | 3.28 (1.03) | 0.0017 |
| rs12249377 | A | *HTR7* | 10 | 90833199 | 4.14 (1.39) | 0.0033 |
| rs4711429 | A | *FKBP5* | 6 | 35715771 | 3.27 (1.12) | 0.0040 |
| rs2420203 | G | *HTR7* | 10 | 90748106 | 3.96 (1.39) | 0.0050 |
| rs11841433 | A | *HTR2A* | 13 | 47434733 | 5.15 (2.26) | 0.0242 |
| rs1328674 | A | *HTR2A* | 13 | 47441707 | 5.15 (2.26) | 0.0242 |
| rs7326071 | A | *HTR2A* | 13 | 47438668 | 5.15 (2.26) | 0.0242 |
| rs3785151 | C | *SLC6A2* | 16 | 55712519 | -3.68 (1.35) | 0.0069 |
| rs2025296 | A | *HTR2A* | 13 | 47463819 | 2.97 (1.67) | 0.0763 |

s.e.: standard error

Table 24. RMSE values and model details of the elastic net analyses.

| Statistic | Full sample (n=319) | LY2216684 (n=157) | Placebo (n=162) |
|---|---|---|---|
| RMSE: baseline covariates | 8.30 | 7.90 | 8.69 |
| RMSE: baseline covariates + SNP | 8.27 | 7.90 | 8.69 |
| Number of SNPs selected | 23 | 0 | 0 |
| α | 0.2 | - | - |
| λ | 4.8 | - | - |

Table 25. SNPs selected by elastic net in full sample analysis.

| SNP | Reference allele | Gene | Chromosome | Position | Parameter estimate |
|---|---|---|---|---|---|
| rs12121815 | G | *KCNK2* | 1 | 215187382 | 0.52 |
| rs4711429 | A | *FKBP5* | 6 | 35715771 | 0.07 |
| rs9377186 | A | *UST* | 6 | 148986631 | 0.25 |
| rs7770997 | G | *UST* | 6 | 149010203 | 0.08 |
| rs556793 | G | *ADRA1A* | 8 | 26841550 | 0.005 |
| rs498917 | A | *ADRA1A* | 8 | 26862186 | 0.31 |
| rs580644 | A | *ADRA1A* | 8 | 26862973 | 0.31 |
| rs505138 | A | *ADRA1A* | 8 | 26869465 | 0.34 |
| rs2420203 | G | *HTR7* | 10 | 90748106 | 0.10 |
| rs12249377 | A | *HTR7* | 10 | 90833199 | 0.47 |
| rs12259062 | G | *HTR7* | 10 | 90850288 | 0.06 |
| rs2070762 | A | *TH* | 11 | 2165105 | 0.35 |
| rs908867 | A | *BDNF* | 11 | 27724217 | -0.001 |
| rs7128320 | G | *GRIK4* | 11 | 120964537 | -0.04 |
| rs12800734 | G | *GRIK4* | 11 | 120966045 | -0.06 |
| rs2276319 | G | *GRIK4* | 11 | 120967112 | -0.10 |
| rs17069005 | G | *HTR2A* | 13 | 46849983 | 0.33 |
| rs17287961 | A | *HTR2A* | 13 | 46862976 | 0.07 |
| rs9316235 | A | *HTR2A* | 13 | 46871568 | 0.13 |
| rs9526245 | C | *HTR2A* | 13 | 46871832 | 0.12 |
| rs731779 | C | *HTR2A* | 13 | 46877903 | 0.08 |
| rs17288723 | G | *HTR2A* | 13 | 46883558 | 0.41 |
| rs174697 | A | *COMT* | 22 | 19966309 | -0.46 |

### 8.3.3. Longitudinal analyses

*8.3.3.1. Outcome variable*

MADRS-TS measurements were available for all patients at week 1, after which the proportion of missing observations increased gradually to 10% at week 10 (Table 26). The mean change in MADRS-TS from baseline decreased over time, with a slope that flattens towards the end of the trial period (Fig.43). Negative outcome values reflect improvements in the patient's condition.

Table 26. Proportion of missing outcome observations.

| Weeks since start of trial | Missing MADRS-TS observations |
|---|---|
| 1 | 0% |
| 3 | 0.3% |
| 5 | 3% |
| 7 | 6% |
| 10 | 10% |

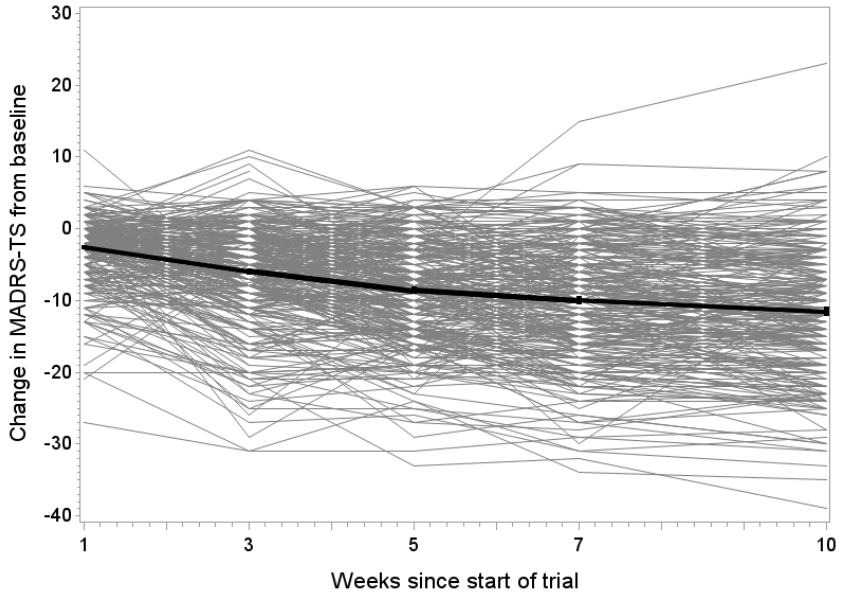



Figure 43. Change in MADRS-TS over trial period per patient (grey) and mean trend (black).

*8.3.3.2. Baseline covariates*

In the longitudinal analysis of baseline covariates, a significant quadratic time effect captured the increase in change in MADRS-TS over time (Table 27). Baseline MADRS-TS and QIDS, as well as the baseline MADRS-TS by time interaction were associated with the outcome variable in all analysis samples. In addition, treatment and the interaction term of treatment by time were significant in the full sample model. Therefore, treatment, baseline MADRS-TS and their interaction terms with time, and baseline QIDS were included as covariates in the longitudinal analyses. Due to the nature of the outcome variable, negative parameter coefficients reflect a positive change in patient health state and vice versa.

All random effects were significant and the unstructured random effects covariance matrix could not be simplified to a more parsimonious structure.

Table 27. Significant covariates in baseline covariates only linear mixed model.

| **Parameter** | **Full sample (n=319)** | | **LY2216684 (n=157)** | | **Placebo (n=162)** | |
|---|---|---|---|---|---|---|
| | Parameter estimate | *p*-value | Parameter estimate | *p*-value | Parameter estimate | *p*-value |
| Time | -0.32 | 0.476 | -1.08 | 0.076 | -0.08 | 0.897 |
| Time² | 0.09 | <.0001 | 0.11 | <.0001 | 0.08 | <.0001 |
| LY2216684 treatment | 1.52 | 0.006 | - | - | - | - |
| LY2216684 treatment by time interaction | -0.54 | <.0001 | - | - | - | - |
| Baseline MADRS-TS | -0.24 | 0.004 | -0.21 | 0.043 | -0.27 | 0.032 |
| Baseline MADRS-TS by time interaction | -0.05 | 0.001 | -0.04 | 0.026 | -0.05 | 0.015 |
| Baseline QIDS | 0.22 | 0.004 | 0.19 | 0.026 | 0.25 | 0.047 |

*8.3.3.3. Linear mixed model analysis*

In the linear mixed model of change in MADRS-TS over time in a SNP-by-SNP analysis, none of the SNPs had a significant effect (Table 28). The significance threshold correcting for the effective number of SNPs tested are $1.31*10^{-4}$ in the full sample, $1.35*10^{-4}$ in the LY2216684 sample and $1.37*10^{-4}$ in the placebo sample.

SNPs in the *HTR2A* gene came up in the top 5 results from each analysis (Table 28). The *HTR2A* SNPs in the full sample and placebo analyses are in mild LD ($r_{LD}^2$>0.65), including one overlapping SNP (rs9316235). The *HTR2A* SNPs from the LY2216684 analysis however are not correlated with the SNPs from other analyses ($r_{LD}^2$<0.03). Furthermore, the two *ADRA1A* SNPS in the LY2216684 model are in perfect LD ($r_{LD}^2$=1) but the *SLC6A2* SNPs from the placebo model are not correlated ($r_{LD}^2$<0.03).

There is similarity in the results of linear regression and linear mixed model analyses in terms of which SNPs have the lowest *p*-values (Fig.44).

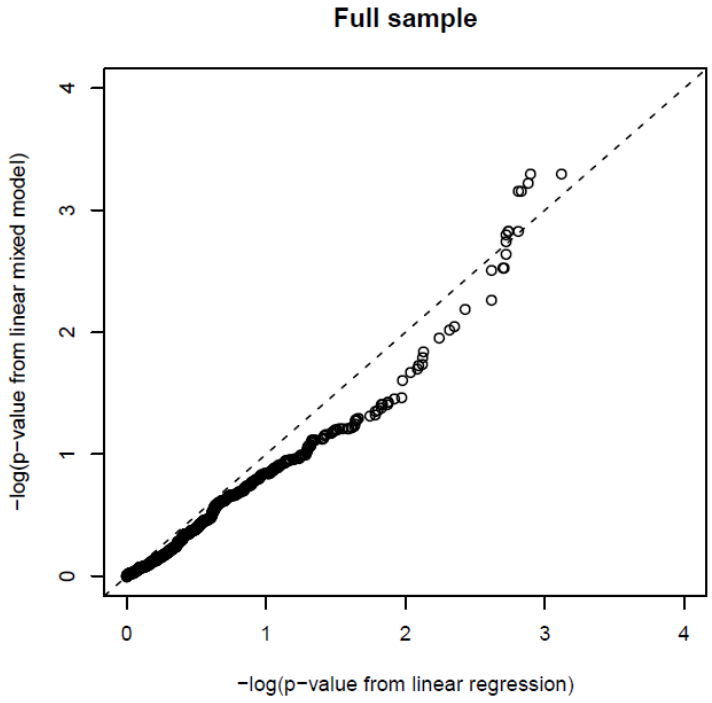



Figure 44. QQ-plots comparing the *p*-values for SNP effect from linear regression analysis with the *p*-values for SNP by time effect from linear mixed model analysis.

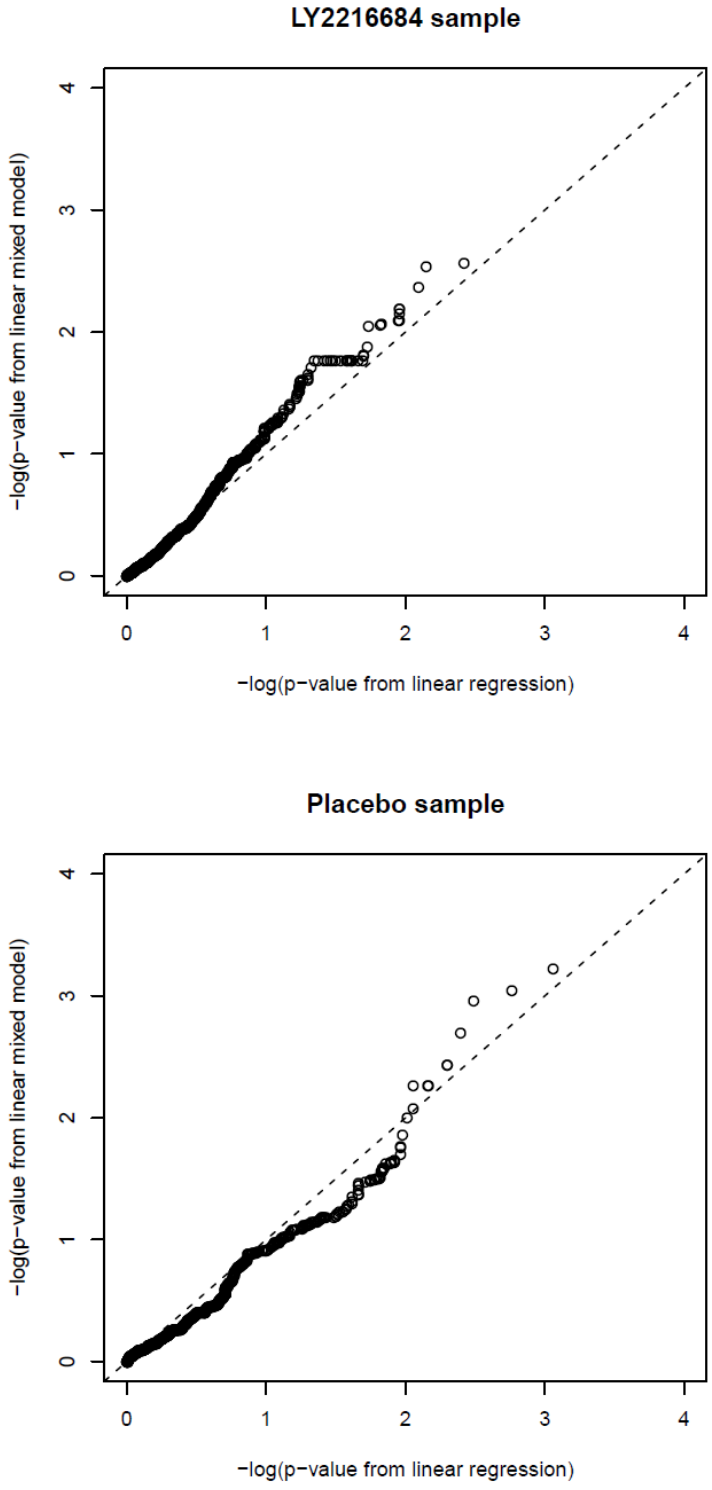



Figure 44 (continued). QQ-plots comparing the *p*-values for SNP effect from linear regression analysis with the *p*-values for SNP by time effect from linear mixed model analysis.

Table 28. SNPs with the five lowest *p*-values for the SNP by time (weeks since start of trial) interaction effect in the full sample and treatment specific linear mixed models. For the LY2216684 and placebo analyses we also report the top 5 SNPs from the full sample model.

| SNP | Reference allele | Gene | Chromosome | Position | Parameter estimate (s.e.) | *p*-value |
|---|---|---|---|---|---|---|
| Full sample analysis (n=319) | | | | | | |
| rs9534501 | A | *HTR2A* | 13 | 46865841 | 0.29 (0.08) | 0.0005 |
| rs9567743 | G | *HTR2A* | 13 | 46866665 | 0.29 (0.08) | 0.0005 |
| rs731779 | C | *HTR2A* | 13 | 46877903 | 0.31 (0.09) | 0.0006 |
| rs9316235 | A | *HTR2A* | 13 | 46871568 | 0.27 (0.08) | 0.0007 |
| rs9526245 | C | *HTR2A* | 13 | 46871832 | 0.27 (0.08) | 0.0007 |
| LY2216684 analysis (n=157) | | | | | | |
| rs12121815 | G | *KCNK2* | 1 | 215187382 | 0.65 (0.21) | 0.0027 |
| rs11841433 | A | *HTR2A* | 13 | 46860598 | -0.46 (0.15) | 0.0029 |
| rs1328674 | A | *HTR2A* | 13 | 46867572 | -0.45 (0.16) | 0.0043 |
| rs498917 | A | *ADRA1A* | 8 | 26862186 | 0.38 (0.14) | 0.0065 |
| rs580644 | A | *ADRA1A* | 8 | 26862973 | 0.38 (0.14) | 0.0065 |
| rs9534501 | A | *HTR2A* | 13 | 46865841 | 0.22 (0.11) | 0.0409 |
| rs9567743 | G | *HTR2A* | 13 | 46866665 | 0.22 (0.11) | 0.0409 |
| rs731779 | C | *HTR2A* | 13 | 46877903 | 0.27 (0.11) | 0.0197 |
| rs9316235 | A | *HTR2A* | 13 | 46871568 | 0.17 (0.10) | 0.0927 |
| rs9526245 | C | *HTR2A* | 13 | 46871832 | 0.17 (0.10) | 0.0927 |
| Placebo analysis (n=162) | | | | | | |
| rs2770298 | G | *HTR2A* | 13 | 46872712 | 0.43 (0.12) | 0.0006 |
| rs747107 | A | *SLC6A2* | 16 | 55661809 | 0.41 (0.12) | 0.0009 |
| rs2770296 | G | *HTR2A* | 13 | 46866425 | 0.41 (0.12) | 0.0011 |
| rs3785151 | C | *SLC6A2* | 16 | 55678607 | -0.43 (0.14) | 0.0020 |
| rs9316235 | A | *HTR2A* | 13 | 46871568 | 0.35 (0.12) | 0.0037 |
| rs9534501 | A | *HTR2A* | 13 | 46865841 | 0.34 (0.12) | 0.0055 |
| rs9567743 | G | *HTR2A* | 13 | 46866665 | 0.34 (0.12) | 0.0055 |
| rs731779 | C | *HTR2A* | 13 | 46877903 | 0.34 (0.14) | 0.0137 |
| rs9316235 | A | *HTR2A* | 13 | 46871568 | 0.35 (0.12) | 0.0037 |
| rs9526245 | C | *HTR2A* | 13 | 46871832 | 0.35 (0.12) | 0.0037 |

s.e.: standard error

*8.3.3.4. Linear mixed elastic net analysis*

The linear mixed elastic net analysis did not select any SNPs that are associated with change in MADRS-TS over time in the full sample. Model convergence was slow: a single cross-

validation iteration for a single tuning parameter combination took from a few hours up to three days to run. Parallelizing the cross-validation steps did speed the process up, but given the relatively small size of the dataset we considered that this was slow. Moreover, the endpoint elastic net with a larger tuning parameter grid could be run in two hours without parallelizing cross-validation. In addition to computational time, we encountered more software weaknesses when estimating the linear mixed elastic net. For some tuning parameter combinations the model failed to converge in some cross-validation folds but did converge in others. The non-convergence occurred across the range of tuning parameter values and was not specific to either low or high values of tuning parameters. An additional issue was the specification of the tuning parameter grid. If the elastic net penalty is too strong (high $\lambda$) all genetic variables are removed from the model, but a penalty that is too weak ($\lambda$ close to zero) retains all SNPs. The range of $\lambda$ that allows variable selection depended heavily on $\alpha$, and construction of the tuning parameter grid required a considerable amount of manual input. For instance, when $\alpha = 1$ values of $\lambda$ smaller than 0.003 retained all genetic variables and $\lambda$ values larger than 0.011 caused all genetic variables to be removed from the model. When $\alpha = 0.6$, the range of $\lambda$ values that allowed variable selection was between 0.005 and 0.025 and thus clearly dependent on the value of $\alpha$. We defined the range of $\lambda$ wherein variable selection can take place by trial and error for each level of $\alpha$, again a time intensive procedure. Because of these computational issues, the analysis was performed on the full sample, but was not repeated in the treatment specific datasets.

#### 8.3.4. Power simulation analysis

We simulated a single SNP for the LY2216684 treated sample, which was analysed as a single SNP using linear regression and in combination with all 1,138 SNPs in an elastic net algorithm. We estimated the power as the proportion of simulations in which the simulated

SNP attained a *p*-value below the multiple-testing threshold ($p= 1.35*10^{-4}$) in linear regression or in which the SNP was retained by variable selection in the elastic net model. The results of this simulation study suggest that in this sample elastic net may have more power than linear regression to detect a single SNP associated with MADRS-TS (Fig.45). Furthermore, the power curve of linear regression is very close to, but slightly lower than the theoretical curve from the genetic power calculator (Purcell, Cherny, & Sham, 2003). For both methods the power when there is no genetic effect, (i.e. heritability equal to zero), is close to zero as expected, indicating that the type I error is well-controlled for in both analysis methods. Realistic heritability values for a single SNP are smaller than 5% and in this range both endpoint methods have low power to detect genetic associations in this sample.

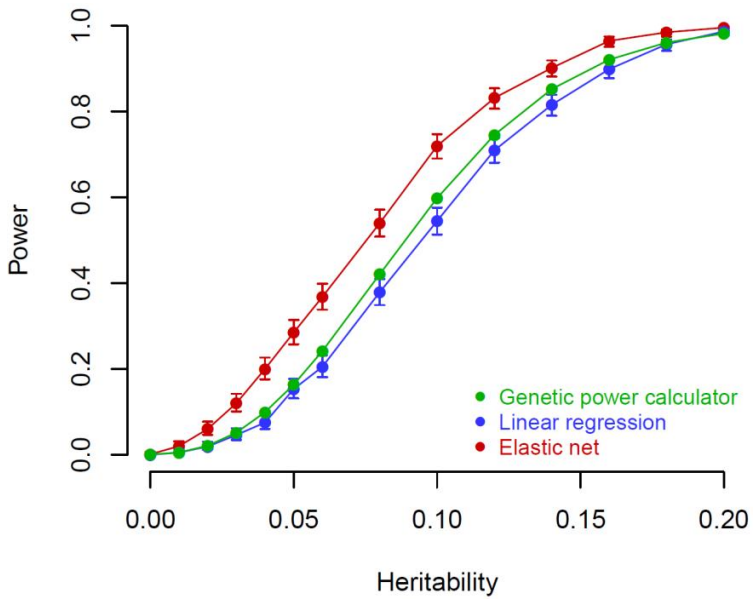



Figure 45. Power curve for linear regression and elastic net, and theoretical power from genetic power calculator (Purcell et al., 2003).

## 8.4. Discussion

This study compared four analysis methods to identify SNPs that have a PGx effect on antidepressant response. We contrasted traditional statistical modelling with machine learning algorithms, both for a single endpoint outcome and a repeated longitudinal outcome.

The elastic net machine learning approach was applied to build an algorithm that predicts change in MADRS-TS from baseline at the end of the trial using a combination of genetic variables. In the full sample, elastic net identified a set of 23 SNPs that increased the predictive performance of the model compared to a clinical baseline covariate model only. Although no SNPs reached significance in the linear regression analysis, the results of both endpoint analyses were consistent in terms of which SNPs provided the strongest evidence for association. The modest decrease in RMSE when SNPs were included in the elastic net implies that the genetic covariates make only a minor contribution to the prediction accuracy. Furthermore, penalized logistic regression has been shown to outperform single SNP analysis methods in detecting associations in a simulation study (Ayers & Cordell, 2010). It is thus not contradictory that these SNPs were picked up by the elastic net but did not reach significance in the linear regression. Nevertheless, our results should be replicated in an independent dataset for confirmation, particularly because our sample size did not allow the data to be split into a separate training and testing subset. In the longitudinal setting we did not detect any SNPs that associated with change in MADRS-TS over time in either the traditional or machine learning analyses.

We have used elastic net to achieve the simultaneous analysis of many SNPs in a single regression-like model for the purpose of variable selection and parameter interpretation, but the resulting model can also be used for outcome prediction. Other machine learning methods, for example support vector machines, may further improve prediction, but these

algorithms are less interpretable. In many machine learning algorithms the relationship between covariates and outcome is not straightforward and difficult to translate to a biological meaning. Elastic net has the advantage that it can be seen as a machine learning extension of linear regression and the interpretation of elastic net model parameters is similar to that of linear regression parameters.

Although statistical learning methods can handle more covariates than subjects, the available sample size is a limitation of our study, as it is not large enough to allow separate training and testing subsets, particularly for the treatment specific samples where $n$ was approximately 160. The average RMSE across cross-validation folds provides an estimate of the predictive performance of the elastic net but is likely to be too optimistic. Nevertheless, this study was primarily interested in the variable selection property of the elastic net and we do not use the model for prediction purposes.

The simulation study revealed that in this sample the elastic net has more power than the linear regression to detect genetic associations with the change in MADRS-TS at the end of the study period. However, as power is defined differently for both analyses, this result should be interpreted with caution. In the linear regression analysis, each SNP is modelled separately and power is the probability of obtaining a significant *p*-value for the simulated SNP. By contrast, elastic net analyses all SNPs in a single model and the probability of the simulated SNP being selected is termed power. There is thus a conceptual difference between the traditional statistical and machine learning approaches. Furthermore, this is a limited simulation study applicable only to our sample and we cannot draw the general conclusion that elastic net is superior to linear regression in all datasets.

Despite the increased power of the elastic net, both methods have less than 30% power to detect genetic effects with 5% heritability in the treatment specific samples. A stronger genetic effect is unlikely, given the lack of success of previous larger studies to identify the

genetic associations with antidepressant response (GENDEP Investigators et al., 2013). Lack of power may therefore be the reason why the elastic net failed to identify any SNPs in the LY2216684 and placebo analyses. Yet, it cannot be ruled out that there simply is no genetic association with the outcome in these samples.

The simulation study also found that the power of the linear regression analysis was slightly lower than that of the theoretically calculated power curve, which is most likely due to the fact that baseline covariates were included in the regression analysis, something the genetic power calculator does not take into account.

Although theoretically feasible, the linear mixed elastic net proved to be computationally challenging. Simplification of the random effects structure did reduce the computational burden significantly, but the software suffered from additional limitations. Non-convergence in some cross-validation folds and difficulties in specifying the tuning parameter grid mean this method is not robust and not practical for applied data analysis. We expect that the computational issues we encountered in our longitudinal analysis will be aggravated if more variables, for example genome-wide data, are analysed.

Other software packages for penalized linear mixed model exist (Table 4), but we did not extend our study to alternative methods for longitudinal statistical learning. The available software packages differ in the parameter estimation algorithm and penalty used, but they are relatively similar. We assume that genetic associations that can be identified by one machine learning method but not by another, are not likely to have clinically or biologically relevant effect sizes.

We conclude that elastic net offers a valuable alternative to traditional statistical methods for the analysis of PGx studies, given its ability to incorporate many covariates in a single model. Elastic net identified 23 SNPs associated with change in MADRS-TS at the study endpoint consistent with linear regression results. Moreover, elastic net seemed to have

more power in our sample than linear regression to detect these associations. The elastic net model can be used for prediction purposes as well as interpretation of the parameter estimates. Although a linear mixed elastic net model is theoretically possible, software needs to be developed further for its application to data analysis.

# 9. Discussion

## 9.1. Summary of thesis research

In this thesis we have studied barriers to the use of PGx testing in clinical practice. We described the characteristics of a clinically useful PGx test for clozapine induced agranulocytosis and noted that the known PGx variants lack sensitivity to alter clozapine monitoring guidelines. We also examined the economic arguments for PGx testing by reviewing the literature and concluded that the majority of economic evaluations found PGx guided pharmacotherapy to be cost-effective. Freely available genetic information would further increase the cost-effectiveness of PGx testing.

Furthermore, we applied machine learning techniques to analyse a gene expression case/control study and two clinical trials with genetic data. These statistical algorithms lend themselves to the analysis of large datasets, such as genome-wide and transcriptome-wide studies. Machine learning methods enable training multivariable models examining large numbers of genetic variables simultaneously, which is a strong advantage over traditional statistical methods. In addition, these algorithms optimize predictive ability and can readily be applied as polygenic PGx tests. Our studies indicated that larger datasets will be necessary to increase the utility of machine learning in PGx research.

## 9.2. Statistical methods for pharmacogenetic research

PGx tests that are currently used in clinical practice are mainly based on a single or a handful of genes with large effect sizes. Most drugs listed on the FDA Table of Pharmacogenomic Biomarkers in Drug Labeling indeed mention a single genetic variant on their label and the maximum number of genes reported per drug is seven (for valproic acid). Single variant successes such as the HLA associations with abacavir and

carbamazepine induced ADRs suggested that genetic associations with PGx outcomes would have large effect sizes. The effect sizes for PGx outcomes detected in GWAS are indeed larger than those for non-PGx complex outcomes, an observation which is not explained by the often smaller sample sizes of PGx studies (Maranville & Cox, 2016). Hence, it was expected that candidate gene studies with small samples would have sufficient power to detect strong PGx associations. Although this approach has been successful for some examples, this has not proved true in general. Many findings from candidate gene studies turned out to be false positives and have failed to replicate, and the success of PGx GWAS in detecting novel PGx variants has been modest (Ioannidis, 2013). Considering the failure to robustly identify strong PGx effects, we need to expand our search to common genetic variants with moderate or weak effect sizes. It is likely that treatment efficacy and safety are complex outcomes and that the genetic contribution is polygenic, i.e. a combined effect of weak and moderate genetic associations. Larger sample sizes will be necessary to identify weak genetic effects. Sample sizes of PGx GWAS have lagged far behind those of disease genetics studies and are often smaller than 1,000 individuals (Motsinger-Reif et al., 2013). In the field of disease genetics, larger sample sizes have indeed led to more genome-wide significant associations and the same can be expected for PGx studies (Gratten et al., 2014). Alternatively, rare variants may have an effect on PGx outcomes through distinct mechanisms. In particular for rare ADRs such as clozapine induced agranulocytosis (chapter 2), this seems a plausible explanation. Investigating rare variant associations requires appropriate study design and statistical methods (Lee, Abecasis, Boehnke, & Lin, 2014).

In addition to larger GWAS to detect modest and weak genetic variants, PGx studies should examine polygenic effects. A polygenic risk score, in effect the sum of several SNP alleles weighted by their effect size in a discovery sample, summarizes genome-wide genetic effects in a single continuous variable (Dudbridge, 2013). This score can subsequently be used to predict phenotypes in a replication sample. For example, polygenic risk scores from

schizophrenia, bipolar disorder and major depressive disorder explain 18.4%, 2.8% and 0.6% of variance in case/control status in these disorders, respectively (Psychiatric GWAS Consortium Bipolar Disorder Working Group, 2011; Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Polygenic risk scores can also be constructed for and used to predict PGx outcomes. Applied to antidepressant response, no significant prediction from polygenic risk scores was obtained, though the authors commented that the study might be underpowered (García-González et al., 2017). As sample sizes increase, the predictive ability of polygenic risk scores may improve and achieve meaningful prediction accuracy for antidepressant response.

As an alternative to polygenic risk scores, machine learning and deep learning methods can be employed to model polygenic effects. Machine learning algorithms allow the simultaneous analysis of large numbers of correlated genetic variants. These methods enable an approach that is not possible via traditional statistical model as the latter struggle with correlated variables. Furthermore, traditional statistical methods such as linear and logistic regression require the number of predictor variables to be small in comparison to the number of individuals in the sample. Machine learning techniques provide a solution to this problem and can even be applied to datasets with more variables than subjects. Moreover, as machine learning algorithms are trained by optimizing prediction accuracy, they can readily be applied as polygenic biomarkers to guide treatment. These methods are also more flexible than polygenic risk scores as machine learning algorithms can model non-linear multivariable effects while polygenic risk scores are linear combinations of single SNP effects. Another advantage is that it is relatively straightforward to include additional predictors such as demographic and clinical variables in machine learning models. In addition, many algorithms return variable importance scores, which can be used to identify relevant genetic variants.

In this thesis, we have applied machine learning to genetic (chapter 7 and 8) and gene expression studies (chapter 6). Both sources of data result in datasets with large numbers of variables, but each type has its own advantages and disadvantages. The invariability of genetic code is an attractive property of genetic data as it means that an individual only needs to be genotyped once to obtain information that remains relevant throughout their life. The same variables, i.e. genotypes, can be used to study many different phenotypes. Moreover, each somatic cell in the body contains the entire genetic code and a simple saliva sample is sufficient to collect it. Although missing genotypes can sometimes be imputed based on LD with observed genotypes, missing observations are inevitable in a genetic dataset. This is not a major problem in GWAS, but many machine learning methods require complete data and thus imputation techniques must be used to estimate missing information. On the other hand, gene expression scores are dynamic over time, so have to be measured at specific time points when used to study treatment response or safety. This means that transcriptomic data are only relevant for a limited period of time. Though the dynamic nature of this type of data also has its advantages as gene expression can change in response to variations in the patient's condition. For example, changes in gene expression following the administration of a drug might predict the development of an ADR. Furthermore, gene expression is tissue and cell specific. Although gene expression in blood cells can easily be measured, other tissues which are less accessible might be more relevant to the outcome studied. For instance, for a PGx study of antipsychotic response gene expression in the brain might be most informative but virtually impossible to obtain. Gene expression can be summarized in a continuous score and by definition there are no missing observations in the data. These characteristics mean that transcriptomic data are particularly suitable for machine learning.

Since machine learning and deep learning methods are capable of analysing large numbers of predictors simultaneously, they are not only convenient for the analysis of genetic or

transcriptomic datasets, but can also be used to build integrative multi-omics prediction models combining several large datasets. When it comes to predicting a PGx outcome, other sources of information than genetics could be relevant. Demographic and clinical descriptives, as well as transcriptomic, proteomic or metabolomic variables may contain some prediction signal and machine learning algorithms can be used to combine predictors from various sources to achieve maximal predictive accuracy. A limitation of this approach is the fact that whereas genetic information is invariable over time and can be stored for future use, other variables have to be measured at the time of testing and results might not be available immediately. For example, it may take days or weeks between sample collection and the reporting of gene expression scores. Nevertheless, combining several sources of information could result in more precise prediction models for treatment efficacy and safety and advance our understanding of the mechanisms underlying PGx outcomes.

## 9.3. Challenges to pharmacogenetic research and implementation

The crucial limitation for successful PGx studies will be obtaining large sample sizes. Not only do participants of PGx efficacy studies need to suffer from the same disease, but they also need to be treated with the same drug. This means the eligible patient population is smaller than for studies of disease genetics, particularly when there is a wide range of treatments available. Most patients treated with a drug achieve average response and contribute little information. Moreover, PGx studies investigating ADRs have more stringent recruitment criteria as they require patients who have developed the ADR. Usually, patients are recruited to a study because they are treated with the specific drug, and few suffer from an ADR. Collecting large samples can prove extremely difficult when the ADR under study is rare.

Secondly, heterogeneity in the phenotype may reduce the power to detect genetic associations and underestimate effect sizes (Manchia et al., 2013). Phenotypic heterogeneity and misdiagnosis is a limitation in disease genetics, but also applies to PGx studies. For example, in a clinical trial evaluating PGx screening prior to abacavir use, *HLA-B*5701* was associated with immunologically confirmed hypersensitivity reaction (odds ratio = 0.03), but the association with clinically diagnosed hypersensitivity reaction was substantially weaker (odds ratio = 0.4) (Mallal et al., 2008). It is thus important to unambiguously define the PGx outcome studied.

It is important to keep in mind that statistical significance is not sufficient for a PGx marker to be of clinical relevance. A genetic association may be significant, but if it does not translate to accurate predictions it lacks clinical utility as a PGx test. In order to make an impact on pharmacotherapy, a PGx test needs to provide reliable results and thus sufficient precision in terms of PPV and NPV. For instance, the *HLA-B*5701* association with abacavir has 100% NPV, meaning that all test negative patients can safely be treated with abacavir. In contrast, a PGx test using a polymorphism in *HLA-DQB1* to predict clozapine induced agranulocytosis identifies one in five agranulocytosis cases and test negative patients benefit from only 20% risk reduction, which is not enough to disregard the risk of this ADR and change blood monitoring requirements. It is unlikely that common variants with weak effect sizes will achieve clinically significant effect sizes on their own. Rather, hope lies in the combination of multiple genetic variants with other information such as clinical variables to achieve higher prediction accuracy. Still, PGx associations with effect sizes below clinical relevance can be useful to elucidate the mechanisms of drug efficacy and safety.

An additional barrier to the implementation of PGx tests is the level of evidence that is required before a PGx effect is considered proven. The highest standard of evidence is a randomized controlled trial, though these are expensive to conduct and time-consuming.

The PGx community needs to discuss the necessity of randomized clinical trials and which other forms of evidence, for example observational studies, can be taken into account (Pirmohamed, 2014). Nevertheless, it is foresighted to incorporate PGx studies in the design of clinical trials or to genotype participants. The pharmaceutical company AstraZeneca plans to sequence the genome of 500,000 trial participants as part of a large genetic project investigating rare genetic variants associated with disease and treatment response (Ledford, 2016).

Since health care budgets are limited, there are economic considerations to be made before a PGx test is used in clinical practice. Our literature review in chapter 4 revealed that over half of economic evaluations found PGx testing was cost-effective. However, one in four economic studies concluded that PGx guided treatment was not cost-effective, which highlights the necessity of assessing the economic consequences of a PGx test prior to implementation in clinical practice. As the cost of genetic testing decreases, it is expected that more PGx tests will be cost-effective.

Once the clinical relevance and economic feasibility of robust PGx associations have been established, PGx testing can be implemented in standard clinical practice. This means that patients outside of the research setting need to be genotyped. It is important that genotyping results are available quickly as not to delay the start of drug treatment. Furthermore, in the case of polygenic prediction algorithms, a large number of genetic variants need to be typed. As PGx testing becomes a more widely used health care application, genome-wide genotypes or even whole genome sequences may be added to electronic health records and genetic information will immediately be available at no additional cost at the time of prescribing.

## 9.4. Future perspectives

Studies with larger sample sizes will be required to improve our chances of identifying novel PGx variants with moderate or weak effect sizes. Not only will larger sample sizes have more power to detect weak effects in GWAS, they will also enhance the potential of machine learning and increase prediction accuracy. Furthermore, advanced deep learning algorithms could be trained on PGx datasets. Since neural networks have outperformed other algorithms in many research areas, they might also excel in predicting PGx outcomes. Given the narrow recruitment criteria for PGx studies it will not be easy to collect large samples. International collaboration can help to achieve this goal. For example, the Clinical Pharmacogenetics Implementation Consortium (CPIC), the International Warfarin Pharmacogenetics Consortium (IWPC) and the Pharmacogenomics Research Network (PGRN) bring together researchers from various countries to join forces on PGx studies (Giacomini et al., 2017).

As well as larger samples, future studies should also collect more information on the study participants. Clinical, demographic, genetic, transcriptomic, epigenetic, proteomic and metabolomic data can be combined to build integrative multi-omics prediction models. Though not a PGx study, the LIBD data analysed in chapter 6 serves as an example of how multiple sources of data can be combined in a single analysis. The LIBD dataset originates from a study investigating differences in post-mortem brains between schizophrenia patients and control subjects. We used gene expression scores from the DLPFC brain region to train a classification algorithm for schizophrenia case/control status. In addition to the DLPFC, gene expression and methylation levels were measured in the hippocampus and caudate and the subjects were genotyped, and these datasets will be released to the scientific community at a later stage of the BrainSeq project (Schubert et al., 2015). Thus, gene expression scores and methylation data from the three brain regions studied could be

combined in a multi-omics machine learning analysis to classify schizophrenia patients from controls. The capacity of machine learning algorithms to include large numbers of variables in a single model enables such integrative analyses where large scale omics datasets are combined in a hypothesis-free way.

The clinical utility of a PGx biomarker depends partly on the predictive precision of the test and its cost-effectiveness. In chapter 2 we have investigated the characteristics that a PGx test for clozapine induced agranulocytosis should have in order to be of clinical relevance. However, there is no widely agreed on threshold for what constitutes a clinically acceptable agranulocytosis risk. A cost-effectiveness study may provide an indication as to the minimum PPV and NPV levels that are required for PGx test to be economically worthwhile. These cost-effective PPV and NPV benchmarks can be used to inform the discussion on clinically relevant prediction accuracy of a PGx test. In addition to health economics, many other clinical and ethical arguments play a role in the debate on clinical implementation of PGx tests. This observation is highlighted by the fact that haematological monitoring of patients on clozapine is not cost-effective when compared to a no monitoring strategy, yet monitoring is compulsory in many countries (Girardin et al., 2014).

Furthermore, it is important that cost-effectiveness assessments are updated regularly. When the input parameters, for example the range of drugs on the market, evolve, the cost-effectiveness of a treatment strategy might change, which is true not only for economic evaluations of PGx tests but for all health care interventions. Newly discovered genetic biomarkers and the price of genetic testing are factors specific to PGx testing that might influence the cost-effectiveness of PGx-guided treatment. We explored the impact of freely available genetic information in chapter 4 and concluded that this variable indeed plays a notable role in the cost-effectiveness of PGx testing. However, it may not always be straightforward to estimate the consequences of changing input parameters on the cost-effectiveness conclusions, in particular as these parameters can evolve concurrently and

can interact with each other. Hence, cost-effectiveness should be re-evaluated when the circumstances of PGx testing change. Up-to-date economic evaluations are necessary to accurately inform the allocation of health care budgets.

In conclusion, to advance the field of PGx larger studies will be necessary. Traditional statistical approaches as well as machine learning and deep learning techniques will be useful to unravel the mechanisms of non-response and ADRs and to build accurate multivariable prediction models. It is likely that polygenic effects, i.e. a combination of multiple genetic variants with weak effect sizes, underlie PGx outcomes. Although weak genetic markers have little predictive power on their own, algorithms combining multiple genetic variants, possibly also including other sources of information, might achieve high prediction accuracy. More accurate predictions will improve the clinical utility and cost-effectiveness of PGx testing and make it a valuable tool for personalized medicine.

# References


Adis International Ltd. (2017). Drug profile: Adomeglivant. Retrieved 11 May, 2017, from http://adisinsight.springer.com/drugs/800031719

Agbedjro, D. (2017). Analysis of longitudinal data in statistical learning. Unpublished oral presentation. King's College London.

Alfirevic, A., & Pirmohamed, M. (2017). Genomics of adverse drug reactions. *Trends in Pharmacological Sciences, 38*(1), 100-109.

Altar, C. A., Carhart, J., Allen, J. D., Hall-Flavin, D., Winner, J., & Dechairo, B. (2015). Clinical utility of combinatorial pharmacogenomics-guided antidepressant therapy: evidence from three clinical studies. *Molecular Neuropsychiatry, 1*(3), 145-155.

Anderson, C. A. (2011). Data Quality Control. In E. Zeggini & A. Morris (Eds.), *Analysis of complex disease association studies* (pp. 95-108). San Diego: Academic Press.

Arranz, M. J., & Kapur, S. (2008). Pharmacogenetics in psychiatry: are we ready for widespread clinical use? *Schizophrenia bulletin, 34*(6), 1130-1144.

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic epidemiology, 34*(8), 879-891.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Neural networks: Tricks of the trade* (pp. 437-478). Berlin Heidelberg, Germany: Springer.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., et al. (2010). *Theano: A CPU and GPU math compiler in Python.* Paper presented at the 9th Python in Science Conference (pp. 1-7).

Biernacka, J., Sangkuhl, K., Jenkins, G., Whaley, R., Barman, P., Batzler, A., et al. (2015). The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. *Translational psychiatry, 5*(4), e553.

Bondell, H. D, Krishna, A., & Ghosh, S. (2009). Function to fit Penalized Mixed Model method of Bondell, Krishna, and Ghosh (2010). Retrieved February 15, 2017, from http://www4.stat.ncsu.edu/~bondell/Software/PenLME/PenLME.R

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics, 66*(4), 1069-1077.

Bousman, C. A., Müller, D. J., Ng, C. H., Byron, K., Berk, M., & Singh, A. B. (2017). Concordance between actual and pharmacogenetic predicted desvenlafaxine dose

needed to achieve remission in major depressive disorder: a 10-week open-label study. *Pharmacogenetics and Genomics, 27*(1), 1.

Bouvy, J. C., De Bruin, M. L., & Koopmanschap, M. A. (2015). Epidemiology of adverse drug reactions in Europe: a review of recent observational studies. *Drug safety, 38*(5), 437-453.

Bray, N. J. (2008). Gene expression in the etiology of schizophrenia. *Schizophrenia bulletin, 34*(3), 412-418.

Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199-231.

Breiman, L. (2001b). Random forests. *Machine Learning, 45*, 5-32.

Cattane, N., Minelli, A., Milanesi, E., Maj, C., Bignotti, S., Bortolomasi, M., et al. (2015). Altered gene expression in schizophrenia: findings from transcriptional signatures in fibroblasts and blood. *PLoS ONE, 10*(2), e0116686.

Chen, C. H., Lee, C. S., Lee, M. T., Ouyang, W. C., Chen, C. C., Chong, M. Y., et al. (2014). Variant GADL1 and response to lithium therapy in bipolar I disorder. *N Engl J Med, 370*(2), 119-128.

Chen, S., Grant, E., Wu, T. T., & Bowman, F. D. (2014). Some recent statistical learning methods for longitudinal high-dimensional data. *Wiley Interdisciplinary Reviews: Computational Statistics, 6*(1), 10-18.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., et al. (2016). *Wide & deep learning for recommender systems.* Paper presented at the 1st Workshop on Deep Learning for Recommender Systems (pp. 7-10).

Chung, W.-H., Hung, S.-I., Hong, H.-S., Hsih, M.-S., Yang, L.-C., Ho, H.-C., et al. (2004). Medical genetics: a marker for Stevens-Johnson syndrome. *Nature, 428*(6982), 486-486.

Collobert, R., & Bengio, S. (2004). *Links between perceptrons, MLPs and SVMs.* Paper presented at the 21st International Conference on Machine Learning.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.

Cosgun, E., Limdi, N. A., & Duarte, C. W. (2011). High-dimensional pharmacogenetic prediction of a continuous trait using machine learning techniques with application to warfarin dose prediction in African Americans. *Bioinformatics, 27*(10), 1384-1389.

Covington, P., Adams, J., & Sargin, E. (2016). *Deep neural networks for youtube recommendations.* Paper presented at the 10th ACM Conference on Recommender Systems.

Daniels, M. A., Kan, C., Willmes, D. M., Ismail, K., Pistrosch, F., Hopkins, D., et al. (2016). Pharmacogenomics in type 2 diabetes: oral antidiabetic drugs. *Pharmacogenomics J, 16*(5), 399-410.

Davies, J. C., Wainwright, C. E., Canny, G. J., Chilvers, M. A., Howenstine, M. S., Munck, A., et al. (2013). Efficacy and safety of ivacaftor in patients aged 6 to 11 years with cystic fibrosis with a G551D mutation. *Am J Respir Crit Care Med, 187*(11), 1219-1225.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., et al. (2013). *Recent advances in deep learning for speech research at Microsoft.* Paper presented at the 2013 IEEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Kaae Sønderby, S., Nouri, D., et al. (2015). Lasagne: First release. Retrieved 3 April, 2017, from http://dx.doi.org/10.5281/zenodo.27878

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78-87.

Drago, A., De Ronchi, D., & Serretti, A. (2009). Pharmacogenetics of antidepressant response: an update. *Human genomics, 3*(3), 257.

Dubé, S., Dellva, M. A., Jones, M., Kielbasa, W., Padich, R., Saha, A., et al. (2010). A study of the effects of LY2216684, a selective norepinephrine reuptake inhibitor, in the treatment of major depression. *Journal of psychiatric research, 44*(6), 356-363.

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet, 9*(3), e1003348.

Dukart, J., Schroeter, M. L., Mueller, K., & Alzheimer's Disease Neuroimaging Initiative. (2011). Age correction in dementia-matching to a healthy brain. *PLoS ONE, 6*(7), e22193.

Electronic Medicines Compendium. (2015). Summary of product characteristics:Tegretol tablets 100mg, 200mg, 400mg. Retrieved 14 March, 2017, from https://www.medicines.org.uk/emc/medicine/1328

Electronic Medicines Compendium. (2016a). Summary of product characteristics: Kalydeco 150 mg film-coated tablets. Retrieved 14 March, 2017, from http://www.medicines.org.uk/emc/medicine/27586/SPC/Kalydeco+150+mg+film-coated+tablets/

Electronic Medicines Compendium. (2016b). Summary of product characteristics: Ziagen 20 mg/ml oral solution. Retrieved 14 March, 2017, from http://www.medicines.org.uk/emc/medicine/2475

Eli Lilly and Company. (2013). Lilly announces edivoxetine did not meet primary endpoint of phase III clinical studies as add-on therapy for major depressive disorder. Retrieved May 15, 2014, from https://investor.lilly.com/releasedetail.cfm?ReleaseID=811751

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115-118.

European Commission. (2008). *Annex 2 of the report on the impact assessment of strengthening and rationalising EU pharmacovigilance.* Retrieved April 7, 2017, from http://ec.europa.eu/health//sites/health/files/files/pharmacos/pharmpack_12_2008/pharmacovigilance-ia-vol2_en.pdf.

Fabbri, C., Crisafulli, C., Calabrò, M., Spina, E., & Serretti, A. (2016). Progress and prospects in pharmacogenetics of antidepressant drugs. *Expert Opinion on Drug Metabolism & Toxicology, 12*(10), 1157-1168.

Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics, 40*(4), 2043.

Fillman, S., Cloonan, N., Catts, V., Miller, L., Wong, J., McCrossin, T., et al. (2013). Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Molecular psychiatry, 18*(2), 206-214.

Fisch, A. S., Perry, C. G., Stephens, S. H., Horenstein, R. B., & Shuldiner, A. R. (2013). Pharmacogenomics of anti-platelet and anti-coagulation therapy. *Curr Cardiol Rep, 15*(7), 381.

Flume, P. A., Liou, T. G., Borowitz, D. S., Li, H., Yen, K., Ordonez, C. L., et al. (2012). Ivacaftor in subjects with cystic fibrosis who are homozygous for the F508del-CFTR mutation. *Chest, 142*(3), 718-724.

Frey, B. (2016). Deep learning meets genome biology. Retrieved May 7, 2017, from https://www.oreilly.com/ideas/deep-learning-meets-genome-biology.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1-22.

García-González, J., Tansey, K. E., Hauser, J., Henigsberg, N., Maier, W., Mors, O., et al. (2017). Pharmacogenetics of antidepressant response: a polygenic approach. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 75*, 128-134.

Garriock, H. A., Kraft, J. B., Shyn, S. I., Peters, E. J., Yokoyama, J. S., Jenkins, G. D., et al. (2010). A genomewide association study of citalopram response in major depressive disorder. *Biological psychiatry, 67*(2), 133-138.

Gejman, P. V., Sanders, A. R., & Duan, J. (2010). The role of genetics in the etiology of schizophrenia. *Psychiatric Clinics of North America, 33*(1), 35-66.

GENDEP Investigators, MARS Investigators, & STAR*D Investigators. (2013). Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. *Am J Psychiatry, 170*(2), 207-217.

Giacomini, K. M., Yee, S. W., Mushiroda, T., Weinshilboum, R. M., Ratain, M. J., & Kubo, M. (2017). Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nature Reviews Drug Discovery, 16*(1), 70-70.

Girardin, F. R., Poncet, A., Blondon, M., Rollason, V., Vernaz, N., Chalandon, Y., et al. (2014). Monitoring white blood cell count in adult patients with schizophrenia who are taking clozapine: a cost-effectiveness analysis. *Lancet Psychiatry, 1*(1), 55-62.

Gratten, J., Wray, N. R., Keller, M. C., & Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature neuroscience, 17*(6), 782-790.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). *Speech recognition with deep recurrent neural networks.* Paper presented at the 2013 IEEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by L 1-penalized estimation. *Statistics and Computing, 24*(2), 137-154.

Grossi, E., Podda, G. M., Pugliano, M., Gabba, S., Verri, A., Carpani, G., et al. (2014). Prediction of optimal warfarin maintenance dose using advanced artificial neural networks. *Pharmacogenomics, 15*(1), 29-37.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157-1182.

Guzman, C. B., Zhang, X. M., Liu, R., Regev, A., Shankar, S., Garhyan, P., et al. (2017). Treatment with LY2409021, a glucagon receptor antagonist, increases liver fat in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*. Advance online publication. doi: 10.1111/dom.12958

Hall-Flavin, D. K., Winner, J., Allen, J., Jordan, J., Nesheim, R., Snyder, K., et al. (2012). Using a pharmacogenomic algorithm to guide the treatment of depression. *Translational psychiatry, 2*(10), e172.

Hall-Flavin, D. K., Winner, J. G., Allen, J. D., Carhart, J. M., Proctor, B., Snyder, K. A., et al. (2013). Utility of integrated pharmacogenomic testing to support the treatment of

major depressive disorder in a psychiatric outpatient setting. *Pharmacogenetics and genomics, 23*(10), 535-548.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* (2nd ed.). New York: Springer.

Hatz, M. H. M., Schremser, K., & Rogowski, W. H. (2014). Is individualized medicine more cost-effective? A systematic review. *PharmacoEconomics, 32*(5), 443-455.

Hetherington, S., Hughes, A. R., Mosteller, M., Shortino, D., Baker, K. L., Spreen, W., et al. (2002). Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. *The Lancet, 359*(9312), 1121-1122.

Hou, J., Seneviratne, C., Su, X., Taylor, J., Johnson, B., Wang, X. Q., et al. (2015). Subgroup identification in personalized treatment of alcohol dependence. *Alcohol Clin Exp Res, 39*(7), 1253-1259.

Hou, L., Heilbronner, U., Degenhardt, F., Adli, M., Akiyama, K., Akula, N., et al. (2016). Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. *Lancet, 387*(10023), 1085-1093.

Hughes, S., Hughes, A., Brothers, C., Spreen, W., & Thorborn, D. (2008). PREDICT‐1 (CNA106030): the first powered, prospective trial of pharmacogenetic screening to reduce drug adverse events. *Pharmaceutical statistics, 7*(2), 121-129.

Hung, S. I., Chung, W. H., Jee, S. H., Chen, W. C., Chang, Y. T., Lee, W. R., et al. (2006). Genetic susceptibility to carbamazepine-induced cutaneous adverse drug reactions. *Pharmacogenet Genomics, 16*(4), 297-306.

Hwang, Y., Kim, J., Shin, J., Kim, J.-I., Seo, J., Webster, M., et al. (2013). Gene expression profiling by mRNA sequencing reveals increased expression of immune/inflammation-related genes in the hippocampus of individuals with schizophrenia. *Translational psychiatry, 3*(10), e321.

Ingelman-Sundberg, M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J, 5*(1), 6-13.

Iniesta, R., Hodgson, K., Malki, K., Maier, W., Rietschel, M., Mors, O., et al. (2015). *Combining clinical and genetic variables to predict antidepressant treatment response: a machine learning approach*. Paper presented at the 23rd World Congress of Psychiatric Genetics (WCPG).

International Diabetes Federation. (2015). *IDF Diabetes Atlas, 7th edn*. Brussels, Belgium: International Diabetes Federation. Retrieved May 30, 2017, from http://www.diabetesatlas.org.

International Warfarin Pharmacogenetics Consortium. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med, 2009*(360), 753-764.

Ioannidis, J. P. (2013). To replicate or not to replicate: the case of pharmacogenetic studies. *Circulation: Cardiovascular Genetics, 6*(4), 413-418.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature genetics, 29*(3), 306-309.

Ising, M., Lucae, S., Binder, E. B., Bettecken, T., Uhr, M., Ripke, S., et al. (2009). A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression. *Archives of general psychiatry, 66*(9), 966-975.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning. With applications in R.* (Vol. 6). New York: Springer.

Ji, Y., Biernacka, J. M., Hebbring, S., Chai, Y., Jenkins, G. D., Batzler, A., et al. (2013). Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. *The pharmacogenomics journal, 13*(5), 456-463.

Jonas, D. E., Evans, J. P., McLeod, H. L., Brode, S., Lange, L. A., Young, M. L., et al. (2013). Impact of genotype-guided dosing on anticoagulation visits for adults starting warfarin: a randomized controlled trial. *Pharmacogenomics, 14*(13), 1593-1603.

Jorgensen, A. L., & Pirmohamed, M. (2011). Risk modeling strategies for pharmacogenetic studies. *Pharmacogenomics, 12*(3), 397-410.

Katsila, T., & Patrinos, G. P. (2015). Whole genome sequencing in pharmacogenomics. *Frontiers in pharmacology, 6*, 61.

Kazda, C. M., Ding, Y., Kelly, R. P., Garhyan, P., Shi, C., Lim, C. N., et al. (2016). Evaluation of efficacy and safety of the glucagon receptor antagonist LY2409021 in patients with type 2 diabetes: 12-and 24-week phase 2 studies. *Diabetes care, 39*(7), 1241-1249.

Kazda, C. M., Frias, J., Foga, I., Cui, X., Guzman, C. B., Garhyan, P., et al. (2017). Treatment with the glucagon receptor antagonist LY2409021 increases ambulatory blood pressure in patients with type 2 diabetes. *Diabetes, Obesity and Metabolism*. Advance online publication. doi: 10.1111/dom.12904

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual review of public health, 34*, 119-138.

Khan, L. M. (2013). Comparative epidemiology of hospital-acquired adverse drug reactions in adults and children and their impact on cost and hospital stay - a systematic review. *European journal of clinical pharmacology, 69*(12), 1985-1996.

Kimmel, S. E., French, B., Kasner, S. E., Johnson, J. A., Anderson, J. L., Gage, B. F., et al. (2013). A pharmacogenetic versus a clinical algorithm for warfarin dosing. *N Engl J Med, 369*(24), 2283-2293.

Kohlrausch, F. B. (2013). Pharmacogenetics in schizophrenia: a review of clozapine studies. *Revista Brasileira de Psiquiatria, 35*, 305-317.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E, 69*(6), 066138.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks.* Paper presented at the Neural Information Processing Systems 2012.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1-26.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Ledford, H. (2016). AstraZeneca launches project to sequence 2 million genomes. *Nature, 532*(7600), 427.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics, 95*(1), 5-23.

Li, M.-X., Yeung, J. M., Cherny, S. S., & Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Human genetics, 131*(5), 747-756.

Li, X., Liu, R., Luo, Z. Y., Yan, H., Huang, W. H., Yin, J. Y., et al. (2015). Comparison of the predictive abilities of pharmacogenetics-based warfarin dosing algorithms using seven mathematical models in Chinese patients. *Pharmacogenomics, 16*(6), 583-590.

Li, X., & Teng, S. (2015). RNA sequencing in schizophrenia. *Bioinformatics and biology insights, 9*(Suppl 1), 53.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18-22.

Lim, K. S., Kwan, P., & Tan, C. T. (2008). Association of HLA-B*1502 allele and carbamazepine-induced severe adverse cutaneous drug reaction among Asians, a review. *Neurology Asia, 13*, 15-21.

Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., et al. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports, 6*. doi: 10.1038/srep26286

Liu, R., Li, X., Zhang, W., & Zhou, H. H. (2015). Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database. *PLoS ONE, 10*(8), e0135784.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*(2), 361-386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, *3*(4), 1710-1737.

Loh, W.-Y. (2017). GUIDE classification and regression trees and forests. Retrieved 5 May, 2017, from http://www.stat.wisc.edu/~loh/guide.html

Lu, M., Lewis, C. M., & Traylor, M. (2017). Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. *bioRxiv*, 098541.

Lubamba, B., Dhooghe, B., Noel, S., & Leal, T. (2012). Cystic fibrosis: insight into CFTR pathophysiology and pharmacotherapy. *Clin Biochem, 45*(15), 1132-1144.

Luzzatto, L., & Seneca, E. (2014). G6PD deficiency: a classic example of pharmacogenetics with on-going clinical implications. *Br J Haematol, 164*(4), 469-480.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic acids research, 39*(suppl 1), D52-D57.

Malhotra, A. K., Correll, C. U., Chowdhury, N. I., Müller, D. J., Gregersen, P. K., Lee, A. T., et al. (2012). Association between common variants near the melanocortin 4 receptor gene and severe antipsychotic drug-induced weight gain. *Archives of general psychiatry, 69*(9), 904-912.

Malhotra, A. K., Zhang, J. P., & Lencz, T. (2012). Pharmacogenetics in psychiatry: translating research into clinical practice. *Molecular psychiatry, 17*(8), 760-769.

Malki, K., Tosto, M. G., Mouriño-Talín, H., Rodríguez-Lorenzo, S., Pain, O., Jumhaboy, I., et al. (2017). Highly polygenic architecture of antidepressant treatment response: Comparative analysis of SSRI and NRI treatment in an animal model of depression. *Am J Med Genet B Neuropsychiatr Genet, 174*(3), 235-250.

Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A., Moore, C., et al. (2002). Association between presence of HLA-B* 5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *The Lancet, 359*(9308), 727-732.

Mallal, S., Phillips, E., Carosi, G., Molina, J.-M., Workman, C., Tomažič, J., et al. (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *New England Journal of Medicine, 358*(6), 568-579.

Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., & Alda, M. (2013). The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS ONE, 8*(10), e76295.

Maranville, J. C., & Cox, N. J. (2016). Pharmacogenomic variants have larger effect sizes than genetic variants associated with other dichotomous complex traits. *Pharmacogenomics J, 16*(4), 388-392.

Maruthur, N. M. (2013). The growing prevalence of type 2 diabetes: increased incidence or improved survival? *Curr Diab Rep, 13*(6), 786-794.

McCormack, M., Alfirevic, A., Bourgeois, S., Farrell, J. J., Kasperavičiūtė, D., Carrington, M., et al. (2011). HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N Engl J Med, 364*(12), 1134-1143.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. *Measurement and evaluation in guidance, 6*(3), 157-163.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., et al. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics, 5*(3), 445-464.

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *The British journal of psychiatry, 134*(4), 382-389.

Motsinger-Reif, A. A., Jorgenson, E., Relling, M. V., Kroetz, D. L., Weinshilboum, R., Cox, N. J., et al. (2013). Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenetics and genomics, 23*(8), 383.

Nielsen, M. A. (2015). *Neural Networks and Deep Learning.* Determination Press. Available at http://neuralnetworksanddeeplearning.com

Ninomiya, Y., & Kawano, S. (2016). AIC for the LASSO in generalized linear models. *Electronic Journal of Statistics, 10*(2), 2537-2560.

Pangallo, B., Dellva, M. A., D'Souza, D. N., Essink, B., Russell, J., & Goldberger, C. (2011). A randomized, double-blind study comparing LY2216684 and placebo in the treatment of major depressive disorder. *J Psychiatr Res, 45*(6), 748-755.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). *Deep face recognition.* Paper presented at the 26th British Machine Vision Conference.

Patel, J. N. (2016). Cancer pharmacogenomics, challenges in implementation, and patient-focused perspectives. *Pharmacogenomics and Personalized Medicine, 9*, 65-77.

Pavani, A., Naushad, S. M., Kumar, R. M., Srinath, M., Malempati, A. R., & Kutala, V. K. (2016). Artificial neural network-based pharmacogenomic algorithm for warfarin dose optimization. *Pharmacogenomics, 17*(2), 121-131.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825-2830.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379.

Pengo, V., Zambon, C.-F., Fogar, P., Padoan, A., Nante, G., Pelloso, M., et al. (2015). A Randomized trial of pharmacogenetic warfarin dosing in naïve patients with non-valvular atrial fibrillation. *PLoS ONE, 10*(12), e0145318.

Perlis, R. H., Smoller, J. W., Ferreira, M. A., McQuillin, A., Bass, N., Lawrence, J., et al. (2009). A genomewide association study of response to lithium for prevention of recurrence in bipolar disorder. *American Journal of Psychiatry, 166*(6), 718-725.

PharmGKB. (2017a). PharmGKB FAQs. Retrieved July 5, 2017, from https://www.pharmgkb.org/page/faqs#what-is-the-difference-between-pharmacogenetics-and-pharmacogenomics

PharmGKB. (2017b). Drug Labels. Retrieved February 8, 2017, from https://www.pharmgkb.org/view/drug-labels.do

Phillips, K. A., Ann Sakowski, J., Trosman, J., Douglas, M. P., Liang, S.-Y., & Neumann, P. (2014). The economic value of personalized medicine tests: what we know and what we need to know. *Genet Med, 16*(3), 251-257.

Pickard, B. S. (2017). Genomics of Lithium Action and Response. *Neurotherapeutics, 14*(3), 582-587.

Pirmohamed, M. (2014). Personalized pharmacogenomics: predicting efficacy and adverse drug reactions. *Annu Rev Genomics Hum Genet, 15*, 349-370.

Pirmohamed, M., Burnside, G., Eriksson, N., Jorgensen, A. L., Toh, C. H., Nicholson, T., et al. (2013). A randomized trial of genotype-guided dosing of warfarin. *N Engl J Med, 369*(24), 2294-2303.

Pirmohamed, M., & Park, B. K. (2001). Genetic susceptibility to adverse drug reactions. *Trends Pharmacol Sci, 22*(6), 298-305.

Pouget, J. G., Shams, T. A., Tiwari, A. K., & Müller, D. J. (2014). Pharmacogenetics and outcome with antipsychotic drugs. *Dialogues in Clinical Neuroscience, 16*(4), 555-566.

Preissner, S. C., Hoffmann, M. F., Preissner, R., Dunkel, M., Gewiess, A., & Preissner, S. (2013). Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. *PLoS ONE, 8*(12), e82562.

Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat Rev Genet, 11*(7), 459-463.

Psychiatric GWAS Consortium Bipolar Disorder Working Group. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature genetics, 43*(10), 977-983.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*(1), 149-150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

R Development Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.5). Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, B. W., Davies, J., McElvaney, N. G., Tullis, E., Bell, S. C., Drevinek, P., et al. (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. *N Engl J Med, 365*(18), 1663-1672.

Reiley, C. (2016). Deep driving. *MIT Rechology Review*. Retreived March 7, 2017, from https://www.technologyreview.com/s/602600/deep-driving/.

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., et al. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry, 18*(4), 497-511.

Roden, D. M., Wilke, R. A., Kroemer, H. K., & Stein, C. M. (2011). Pharmacogenomics: the genetics of variable drug responses. *Circulation, 123*(15), 1661-1670.

Rohart, F. (2011). Multiple hypotheses testing for variable selection. *arXiv preprint arXiv:1106.3415*.

Rohart, F. (2016). Multiple hypothesis testing for variable selection. *Australian & New Zealand Journal of Statistics, 58*(2), 245-267.

Rohart, F., San Cristobal, M., & Laurent, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Computational Statistics & Data Analysis, 80*(0), 209-222.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Sainz, J., Mata, I., Barrera, J., Pérez-Iglesias, R., Varela, I., Arranz, M. J., et al. (2013). Inflammatory and immune response genes have significantly altered expression in schizophrenia. *Molecular psychiatry, 18*(10), 1056.

Samwald, M. (2017). Medication safety code initiative. Retrieved 24 April, 2017, from http://safety-code.org/providers/

Schelldorfer, J., Bühlmann, P., & Van De Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using $\ell_1$-penalization. *Scandinavian Journal of Statistics, 38*(2), 197-214.

Schelldorfer, J., Meier, L., & Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using $\ell_1$-penalization. *Journal of Computational and Graphical Statistics, 23*(2), 460-477.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*(7510), 421-427.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Schubert, C. R., O'Donnell, P., Quan, J., Wendland, J. R., Xi, H. S., Winslow, A. R., et al. (2015). BrainSeq: neurogenomics to drive novel target discovery for neuropsychiatric disorders. *Neuron, 88*(6), 1078-1083.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221-242.

Sharabiani, A., Bress, A., Douzali, E., & Darabi, H. (2015). Revisiting warfarin dosing using machine learning techniques. *Computational and mathematical methods in medicine*. doi:10.1155/2015/560108

Singh, A. B. (2015). Improved antidepressant remission in major depression via a pharmacokinetic pathway polygene pharmacogenetic report. *Clinical Psychopharmacology and Neuroscience, 13*(2), 150-156.

Song, J., Bergen, S. E., Di Florio, A., Karlsson, R., Charney, A., Ruderfer, D. M., et al. (2017). Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder. *Mol Psychiatry, 22*(8), 1223.

Song, J., Bergen, S. E., Di Florio, A., Karlsson, R., Charney, A., Ruderfer, D. M., et al. (2016). Genome-wide association study identifies SESTD1 as a novel risk gene for lithium-responsive bipolar disorder. *Mol Psychiatry, 21*(9), 1290-1297.

Spear, B. B., Heath-Chiozzi, M., & Huff, J. (2001). Clinical application of pharmacogenetics. *Trends Mol Med, 7*(5), 201-204.

Spina, E., & de Leon, J. (2015). Clinical applications of CYP genotyping in psychiatry. *Journal of Neural Transmission, 122*(1), 5-28.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(1), 1929-1958.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44-47.

Sultana, J., Cutroneo, P., & Trifirò, G. (2013). Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology and Pharmacotherapeutics, 4*(5), 73.

Swen, J., Nijenhuis, M., Boer, A. d., Grandia, L., Maitland-van der Zee, A.-H., Mulder, H., et al. (2011). Pharmacogenetics: from bench to byte - an update of guidelines. *Clinical Pharmacology & Therapeutics, 89*(5), 662-673.

Symonds, W., Cutrell, A., Edwards, M., Steel, H., Spreen, B., Powell, G., et al. (2002). Risk factor analysis of hypersensitivity reactions to abacavir. *Clinical therapeutics, 24*(4), 565-573.

Taber, K. A. J., & Dickinson, B. D. (2015). Genomic-based tools for the risk assessment, management, and prevention of type 2 diabetes. *The application of clinical genetics, 8*, 1-8.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). *Deepface: Closing the gap to human-level performance in face verification.* Paper presented at IEEE Conference on Computer Vision and Pattern Recognition 2014.

Tang, J., Liu, R., Zhang, Y.-L., Liu, M.-Z., Hu, Y.-F., Shao, M.-J., et al. (2017). Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients. *Scientific Reports, 7*. doi: 10.1038/srep42192

Tansey, K. E., Guipponi, M., Hu, X., Domenici, E., Lewis, G., Malafosse, A., et al. (2013). Contribution of common genetic variants to antidepressant response. *Biol Psychiatry, 73*(7), 679-682.

U.S. Food and Drug Administration. (2013a). Amaryl prescribing information. Retrieved March 30, 2017, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020496s027lbl.pdf

U.S. Food and Drug Administration. (2013b). Carbatrol prescribing information. Retrieved October 5, 2016, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/020712s032s035lbl.pdf

U.S. Food and Drug Administration. (2015). Ziagen prescribing information. Retrieved October 5, 2016, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2015/020977s030,020978s034lbl.pdf

U.S. Food and Drug Administration. (2016a). Coumadin prescribing information. Retrieved October 5, 2016, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2016/009218s116lbl.pdf

U.S. Food and Drug Administration. (2016b). Glucotrol prescribing information. Retrieved 30 March, 2017, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2016/017783s026lbl.pdf

U.S. Food and Drug Administration. (2017a). Kalydeco prescribing information. Retrieved October 5, 2016, from http://www.accessdata.fda.gov/drugsatfda_docs/label/2017/203188s022l_207925s003lbl.pdf

U.S. Food and Drug Administration. (2017b). Table of pharmacogenomic biomarkers in drug labeling. Retrieved February 8, 2017, from http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

Uher, R., Perroud, N., Ng, M. Y., Hauser, J., Henigsberg, N., Maier, W., et al. (2010). Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. *Am J Psychiatry, 167*(5), 555-564.

Umeda-Yano, S., Hashimoto, R., Yamamori, H., Weickert, C. S., Yasuda, Y., Ohi, K., et al. (2014). Expression analysis of the genes identified in GWAS of the postmortem brain tissues from patients with schizophrenia. *Neuroscience letters, 568*, 12-16.

Van Driest, S. L., Shi, Y., Bowton, E. A., Schildcrout, J. S., Peterson, J. F., Pulley, J., et al. (2014). Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clin Pharmacol Ther, 95*(4), 423-431.

VanVoorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology, 3*(2), 43-50.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Verbelen, M., Collier, D. A., Cohen, D., MacCabe, J. H., & Lewis, C. M. (2015). Establishing the characteristics of an effective pharmacogenetic test for clozapine-induced agranulocytosis. *Pharmacogenomics J, 15*(5), 461-466.

Verbelen, M., & Lewis, C. M. (2015). How close are we to a pharmacogenomic test for clozapine-induced agranulocytosis? *Pharmacogenomics, 16*(9), 915-917.

Verbelen, M., Weale, M. E., & Lewis, C. M. (2017). Cost-effectiveness of pharmacogenetic-guided treatment: are we there yet? *Pharmacogenomics J*. Advance online publication. doi: 10.1038/tpj.2017.21

Verhoef, T. I., Ragia, G., de Boer, A., Barallon, R., Kolovou, G., Kolovou, V., et al. (2013). A randomized trial of genotype-guided dosing of acenocoumarol and phenprocoumon. *N Engl J Med, 369*(24), 2304-2312.

Weale, M. E. (2010). Quality control for genome-wide association studies. In M. R. Barnes, & G. Breen (Eds.), *Genetic Variation: Methods and Protocols*, (341-372). Totowa, New York: Humana Press.

Wetterstrand, K. A. (2016). DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP). Retrieved January 27, 2016, from www.genome.gov/sequencingcosts

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., et al. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther, 92*(4), 414-417.

Winner, J. G., Carhart, J. M., Altar, A., Allen, J. D., & Dechairo, B. M. (2013). A prospective, randomized, double-blind study assessing the clinical impact of integrated pharmacogenomic testing for major depressive disorder. *Discovery medicine, 16*(89), 219-227.

World Health Organization. (2016). *Global report on diabetes*. Geneva: World Health Organization. Available from http://www.who.int/diabetes/global-report/en/

Zhou, K., Pedersen, H. K., Dawed, A. Y., & Pearson, E. R. (2016). Pharmacogenomics in diabetes mellitus: insights into drug action and drug discovery. *Nat Rev Endocrinol, 12*(6), 337-346.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*(5), 654-657.